A Unified Scientific Basis for Inference.

Inge S. Helland
University of Oslo, Oslo, Norway
ingeh@math.uio.no

Abstract

Every experiment or observational study is made in a context. This context is being explicitly considered in this book. To do so, a conceptual variable is defined as any variable which can be defined by (a group of) researchers in a given setting. Such variables are classified. Sufficiency and ancillarity are defined conditionally on the context. The conditionality principle, the sufficiency principle and the likelihood principle are generalized, and a tentative rule for when one should not condition on an ancillary is motivated by examples. The theory is illustrated by the case where a nuisance parameter is a part of the context, and for this case, model reduction is motivated. Model reduction is discussed in general from the point of view that there exists a mathematical group acting upon the parameter space. It is shown that a natural extension of this discussion also gives a conceptual basis from which essential parts of the formalism of quantum mechanics can be derived. This implies an epistemological basis for quantum theory, a kind of basis that has also been advocated by part of the quantum foundation community in recent years. Born's celebrated formula is shown to follow from a focused version of the likelihood principle together with some reasonable assumptions on rationality connected to experimental evidence. Some statistical consequences of Born's formula are sketched. The questions around Bell's inequality are approached by using the conditionality principle for each observer. The objective aspects of the world are identified with the ideal inference results upon which all observers agree (epistemological objectivity).

1 Introduction.

The aim of science is to gain knowledge about the external world; this is what we mean by an epistemic process. In its most primitive form, the process of achieving knowledge can be described by what Brody (1993) called an epistemic cycle: "Act, and see what happens". Experiments in laboratories and observational studies done by scientists are usually much more sophisticated than this; they often require several epidemic cycles and also higher order epistemic cycles acting upon the first order cycles. An experiment or an observational study is always focused on some concrete system, it involves concrete experimental/observational

questions and it is always done in a context, which might depend on conceptual formulations; in addition the context may be partly historical and partly chosen by the scientist himself, or depending upon the scientist.

In earlier years, experiments were often done by single scientists; now it is more and more common that people are working in teams. Also, results of experiments should be communicated to many people. This calls for a conceptual basis which is common to a whole culture of scientists. One problem, however, is that people from different scientific cultures have difficult with communicating. They might not have a common language. The first purpose of this book is to develop a scientific language for achieving knowledge which is a synthesis of the languages that I have met in the three cultures that have been exposed to myself: 1) Mathematical statistics; 2) Quantum mechanics; 3) Applied statistics including simple applications and also to some extent chemometrics. It is a hope that this investigation may lead to a deeper understanding of the epistemic process itself, and thus perhaps imply an enrichment of these different cultures. It is also a hope that such an investigation may be continued in order to include more scientific cultures, say, official statistics, machine learning and quantum computation.

Since statistics is used as a tool in very many experimental studies, also within physics, it is natural to take this culture as a point of departure. But I will add some elements which are not very common in the statistical literature:

1. I make explicit that every experimental investigation is made in a context.

2. A transformation group may be added to the statistical model.

3. Model reductions by means of such groups are introduced.

4. In order to be more general, the parameter concept is replaced by that of an epistemic conceptual variable (e-variable). This notion may also include latent variables, and an e-variable can also be connected to a single unit (say to a single human being in a sociological or psychological investigation or to a single particle in physics). The basic aim of an epistemic process is to gain some knowledge about the relevant e-variables.

5. To find a conceptual epistemic basis for quantum mechanics, I will also introduce inaccessible e-variables, that is, conceptual variables which cannot be estimated with arbitrary accuracy in any experiment. Macroscopic versions of such unknown variables can be found in counterfactual situations, but the notion is also relevant, say, in connection to regression models where the number of variables by necessity is larger than the number of units.

Also, I have included the recent notion of confidence distributions, in order to allow both a frequentist and a Bayesian basis for any given experimental investigation.

It is crucial that this framework as further developed in the present book leads to a non-formal basis for essential elements of quantum theory, a theme

which occupies the second part of the book. The now traditional formal basis for quantum theory as developed by von Neumann (1932) was a great achievement, and the language that is implied by this basis has been used for all further theoretical developments and for all discussions among physicists since then. It is a strong intrinsic part of the quantum mechanical culture, in fact, of the culture shared by the whole community of modern physicists. On this background it may seem presumptuous to claim this, but I will say it anyway: In my view, the development of an alternative to this language is long overdue. The traditional language is purely formal and has little or no intuitive basis for people outside the community of physicists and mathematicians.

Many recent investigators in quantum foundations have reasoned that quantum mechanics should be interpreted as an epistemic science. I agree with this. But I see it as problematic that this notion of an epistemic science should be connected to one language in fundamental physics and a completely different language in the rest of empirical science. The purpose of this book is to work out a common language with a simple intuitive basis. I offer translations of this to the traditional quantum physical language. There are open ends of the present program as far as quantum physics is concerned, but I will argue at the end that the investigations can be carried on further along the same lines.

The sceptic might ask: What is the purpose of introducing a new language when this does not lead to anything new? My answer is that I will show that my program indeed leads to something essentially new, also within the science of quantum mechanics itself: The Born formula, which is the basis for all probability calculations in quantum physics, is taken as an independent axiom in the traditional formulation. I will derive it from a set of intuitive assumptions. In my opinion I also resolve the problematic questions connected to Bell's inequalities by using statistical principles. Also questions around the derivation of the Schrödinger equation are discussed.

The notion of scientific cultures is interesting. In my view it can be seen in the same setting as human cultures in general. Every human being has a background in some culture. He/she has a unique personal history, and this history, together with his/her free will, determines his/her actions at any point of time. People with a similar history often group together and develop cultures. Today we see that large parts of the world are becoming multicultural, and it is important that we achieve understanding across the different cultures. This calls upon rational behavior and rational upbringing in all settings. In this context it is important that science, as potentially the most rational activity among human beings, can be given a basis which is as universal as possible.

The study of scientific cultures is not common. An exception is the book by Knorr Cetina (1999), where the author describes from the inside epistemic cultures connected to two empirical groups: High energy physics experimenters at CERN and molecular biologists at a laboratory. Her arguments strongly depend upon the notion of knowledge societies. Of course I agree that the nature of knowledge is different in different scientific communities, but it is the process of *achieving knowledge* that I feel should have something in common, and it is this process I will focus upon in this book.

PART I

2 The statistical culture.

2.1 Probability

.

The concept of probability will be taken as the same both in statistics and in quantum physics. As developed by Kolmogorov in 1930, it is a taken as a normed measure on some probability space Ω.

Formally, we first introduce a σ-algebra $\mathcal{F}$ of subsets of Ω. The mathematical requirements are: a) Ω should be in $\mathcal{F}$; b) the complement A^c should be in $\mathcal{F}$ whenever A is in $\mathcal{F}$, where $A^c = \{\omega \in \Omega : \omega \notin A\}$; c) $\cup_{n=1}^{\infty} A_n$ should be in $\mathcal{F}$ whenever A_n is in $\mathcal{F}$ for $n = 1, 2, ...$

A normed measure P is then a set function such that a) $0 \leq P(A) \leq P(\Omega) = 1$ for all $A \in \mathcal{F}$; b) $P(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$ if the sets A_i and A_j are disjoint; i.e., $A_i \cap A_j = \emptyset$ when $i \neq j$. This implies $P(A^c) = 1 - P(A)$ and $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ for all sets A and B. The sets $A \in \mathcal{F}$ are called events, and the triple $(\Omega, \mathcal{F}, P)$ is called a probability space.

If Ω is a topological space, the Borel σ-algebra is the smallest σ-algebra containing all open sets. If Ω is discrete and finite, we can, and will, take $\mathcal{F}$ to consist of all subsets.

A random variable X is a measurable function from Ω into the Euclidean space $\mathcal{R}^n$, that is, a function such that $\{X \in B\} = \{\omega \in \Omega : X(\omega) \in B\} = X^{-1}(B)$ is in $\mathcal{F}$ whenever B is a Borel set in $\mathcal{R}^n$. The probability distribution of X is defined by $P(X \in B) = P(X^{-1}(B))$.

Readers not willing to go into all these mathematical details may think of a random variable X as some variable with a distribution associated with it. In this book I will work with real-valued random variables of two kinds:

- Discrete finite-valued random variables X with point probabilities $p(i) = P(X = i); i = 1, ..., r$ satisfying $\sum_{i=1}^{r} p(i) = 1$.
- Continuous random variables X with $P(a \leq X \leq b) = \int_a^b f(x)dx$ for some probability density $f(x)$ satisfying $\int_{-\infty}^{\infty} f(x)dx = 1$.

From this we define expectation

$$\mu = \mathrm{E}(X) = \sum_{i=1}^{r} ip(i); \quad \mu = \mathrm{E}(X) = \int_{-\infty}^{\infty} xf(x)dx$$

$$\mathrm{E}(g(X)) = \sum_{i=1}^{r} g(i)p(i); \quad \mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$$

and variance

$$\sigma^2 = \mathrm{Var}(X) = E[(X - \mu)^2].$$

Two random variables X and Y are independent if $P((X \in A) \cap (Y \in B)) = P(X \in A)P(Y \in B)$ for all Borel sets A and B, with a natural generalization to several random variables. For discrete random variables, this is equivalent to $p_{X,Y}(i,j) = p_X(i)p_Y(j)$; for continuous random variables it is equivalent to $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ with an obvious definition of the joint density.

The interpretation of the probability concept is important for applications. In the literature, three different, but related interpretations are given:

1. The principle of equally likely outcome: If there are r possible outcomes, each is given the probability $1/r$. This can immediately be applied to discrete finite-valued variables, and has examples in the tossing of a die, the tossing of a coin, in card games, in opinion polls etc. Below I will generalize the principle to random variables with a compact range, using the group concept. When the range is not compact, we need un-normed measures. This causes conceptual difficulties that I will not go too deeply into in the present book. I will return to the problem at some points, however.

2. The principle of odds making or subjective probability: The probability of an event A is found on the basis of how much a person is willing to pay for each outcome in a wager with the two outcomes A and A^c. This was introduced by de Finetti and Savage, and used by them as a foundation for Bayesian statistics.

3. The principle of long run frequency: If an experiment is repeated n times, the relative frequency of the event A is the number of times A happens, divided upon n. The probability of A is interpreted as the limit in some sense (see below) of the relative frequency as $n \to \infty$. I will indicate below that this interpretation always can be applied, and made precise, in situations where an experiment can be repeated an arbitrary number of times and the probability can be defined from other considerations..

In many concrete applications, not only one, but two or three of these interpretations may be relevant.

The concept of conditional probability can be given a precise mathematical definition using the notion of a Radon-Nikodym derivative. Specifically, if $\mathcal{B}$ is a sub-σ-algebra of $\mathcal{F}$, then we define $P(A|\mathcal{B})$ as the unique (up to a P-measure 0) $\mathcal{B}$-measurable function such that

$$\int_B P(A|\mathcal{B})P(d\omega) = \int_B I_A(\omega)P(d\omega) \tag{1}$$

for all $B \in \mathcal{B}$. If $\mathcal{B}$ is generated by the disjoint events $\{B_i\}$, this is consistent with the definition that $P(A|B_i) = P(A \cap B_i)/P(B_i)$ whenever $P(B_i) \neq 0$.

Finally, asymptotic considerations may simplify statistics in cases where there are many observations. I will introduce three limit concepts in probability:

- Convergence in probability: $P(|Y_n - Y| > \epsilon) \to 0$ for all $\epsilon > 0$.

- Convergence almost surely: $P(\{\omega : Y_n(\omega) \to Y(\omega)\}) = 1$. This is stronger than convergence in probability.
- Convergence in law: $P(Y_n \leq y) \to P(Y \leq y)$ for all y where $F(y) = P(Y \leq y)$ is continuous. This is a property of the distribution functions rather than of the random variables, but it is related in several ways to the concept of convergence in probability. For instance, for convergence to a degenerate distribution, the two concept are equivalent, and when Y_n tends in law to Y and U_n in probability to c, then $Y_n + U_n$ tends in law to $Y + c$.

In statistical applications, we often have repeated observations, and thus a sample $\boldsymbol{X} = (X_1, X_2, ..., X_n)$, where the X_i's are independent with the same distribution. Assuming these have finite expectation μ and finite variance σ^2, one can prove three limit laws for the mean $\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i$:

- The law of large numbers I: $\bar{X}_n$ converges in probability to μ as $n \to \infty$.
- The law of large numbers II: $\bar{X}_n$ converges almost surely to μ as $n \to \infty$.
- The central limit theorem: $\sqrt{n}(\bar{X}_n - \mu)$ converges in law to $N(0, \sigma^2)$, where $N(\xi, \sigma^2)$ in general is the continuous distribution with density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\xi)^2}{2\sigma^2}). \tag{2}$$

For the first and the third law, see Lehmann (1999). The second law is proved for instance in Sen and Singer (1993).

Using the law of large numbers on the indicator functions $X_i = I(Z_i \in A)$, one easily shows that the frequency interpretation of the probability concept always is valid in situations where it is applicable.

2.2 Statistical models.

In general, a model is a representation of the real world, simplified, but designed such that the essential features that one is interested in, are focused in the model and are correctly represented in the model. A map of the London underground is sometimes taken as an example of a model.

In statistics, one wants a model which can be employed in the epistemic process. This is the reasoning used: The unknown feature that one is interested in, is modeled as a *parameter* θ, real-valued or belonging to a subset of some Euclidean space $\mathcal{R}^n$. (I will not go into nonparametric statistics in this book.) Giving θ some value defines a state of the unknown world. Look at the situation before the experiment or observational study is done, and choose some potential *observations* X_i. These observations are assumed to have a probability distribution for each given state of the world. *The specification of this class of probability distributions constitutes the statistical model.* The statistical model should focus on the relationship between the parameter that one is interested

in and the observations to be done, and it should represent this relationship as well as possible. It should be simple, but not too simple.

The purpose of statistical modeling can be listed as follows:

- Give a rough description of the data generating process.
- Provide parameters that can be estimated from data.
- Allow focusing upon certain parameters.
- Give a language for asking questions about nature.
- Give means for answering such questions by estimation or by the testing of hypotheses.
- Provide confidence intervals and error estimates.
- Give a possibility to study deviations from the model and choosing new models.

Thus the model can be seen as part of a language. A model should be chosen carefully using subject matter knowledge together with a realization of what can be done statistically. If possible, the model should be scrutinized empirically. But once the model is chosen, this is an existential choice. Any conclusion is conditional, given the model.

Specifically, consider the situation with n repeated, independent observations. This is modeled by independent, identically distributed random variables $(X_1, X_2, ..., X_n)$ with distribution depending upon some parameter θ,

In the discrete case:

$$P_\theta(X_1 = x_1, ..., X_n = x_n) = \Pi_{k=1}^n p_\theta(x_k).$$

In the continuous case:

$$P_\theta(X_1 \leq x_1, ..., X_n \leq x_n) = \int_{-\infty}^{x_1} ... \int_{\infty}^{x_n} \Pi_{k=1}^n f_\theta(u_k) du_1 ... du_n.$$

Here $p_\theta(x)$ is the point probability of the individual observations and $f_\theta(x)$ is the probability density of the individual observations. For continuous models, a very common choice of the probability density $f_\theta(x)$ is the normal density (2). In some cases, this may be motivated by some form of the central limit theorem; in other cases it is just a matter of convenience. Here $\theta = (\xi, \sigma)$. One can distinguish between three cases: 1) σ is known and ξ is the unknown parameter. 2) ξ is known and σ is the unknown parameter. 3) Both ξ and σ are unknown. The study of statistical methods that are robust against the assumption of normality, is an active research area in statistics.

The simplest discrete case is when one has n independent repeated trials, each with two possible outcomes A or A^c, often called success and failure. Assuming the same unknown probability θ of success in each trial, and letting

X_k be the indicator of success in the kth trial, we have the point probabilities $p_\theta(0) = 1-\theta$ and $p_\theta(1) = \theta$. If now Y is the number of successes in the n trials, it is a straightforward exercise to show that Y has a binomial distribution:

$$P_\theta(Y = y) = \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y}.$$

Other standard choices of point probabilities and probability densities are listed in Chapter 1 of Lehmann (1999).

2.3 Inference for continuous parameters.

The modern theory of statistical inference was developed by R.A. Fisher in the 1920's and the 1930's, at the same time as modern quantum theory was developed. Fisher knew about quantum theory, but did never hint at any relation to it in his own work.

From an epistemic point of view it is important in statistics to distinguish between the situation before any observations are done, and after observations are done. Before, the observations are unknown, but are modeled as stochastic variables $\boldsymbol{X}$ through the chosen statistical model. After the observations, they are known values $\boldsymbol{X} = \boldsymbol{x}$, and we want to use these observations to say something about the state of nature, the parameters θ. There has to be a recipe from $\boldsymbol{x}$ to the inference about θ.

The simplest concept is that of point estimation: The parameter θ is estimated by a function of the data: $\widehat{\theta}(\boldsymbol{x})$. The properties of this estimation procedure is evaluated by looking at the before-observation situation and using the statistical model: With the stochastic variable $\boldsymbol{X}$ inserted, $\widehat{\theta}(\boldsymbol{X})$ is called an *estimator*. One good property might be that the estimator is unbiased: $\mathrm{E}(\widehat{\theta}(\boldsymbol{X})) = \theta$ or nearly so. Another good property is that it has a small variance. These two properties are sometimes combined in the requirement that the estimator should have a mean square error which is as small as possible, where

$$MSE(\widehat{\theta}(\boldsymbol{X})) = \mathrm{E}((\widehat{\theta}(\boldsymbol{X}) - \theta)^2) = \mathrm{Var}(\widehat{\theta}(\boldsymbol{X})) + (\mathrm{E}(\widehat{\theta}(\boldsymbol{X})) - \theta)^2.$$

A point estimator is often given together with a *standard error*: An estimate of the standard deviation of the corresponding estimator, i.e., the square root of its variance. The standard error gives an indication of uncertainty of the estimate.

In a typical before-observations situation, one has also the possibility to decide how much data one should take; this may be indexed by a number n. The simplest, but not uncommon, case is that of repeated measurements, that is, of n independent, identically distributed observations $\boldsymbol{X}_n = (X_1, ..., X_n)$, but many more situations of this kind exist. The before-observation version of the estimation recipe, $\widehat{\theta} = \widehat{\theta}(\boldsymbol{X}_n)$ is then the estimator. A weak, but desirable property of the estimator is that it should be *consistent*: $\widehat{\theta}(\boldsymbol{X}_n)$ should converge in probability or almost surely to θ as n tends to infinity. A further property which is often satisfied by some central limit type theorem is that of asymptotic

normality: Typically $\sqrt{n}(\widehat{\theta}-\theta)$ converges in law to $N(0,\sigma^2)$ for some variance σ^2. It is desirable that σ^2 should be as small as possible.

In fact these properties are satisfied for a large class of situations, explored by Fisher. In the case of discrete observations, the model implies joint probabilities for the data $p_\theta(\boldsymbol{x})$, in the continuous case joint probability densities $f_\theta(\boldsymbol{x})$. In both cases, when one focuses upon the θ-dependence, this function is called the likelihood $L(\theta)$. Fisher argued that one should maximize the likelihood to find good estimates of θ. The intuitive argument is that this will provide the θ-value which in a best possible way explains the obtained data. The maximum $\widehat{\theta}(\boldsymbol{x})$ is called the maximum likelihood estimate. Local extremes can be found by equating the derivative of the likelihood functions to zero. In more complicated situations one may have problems with several local extremes, but often these problems may be tackled by numerical maximization methods.

Maximum likelihood estimation is used throughout statistics in a large number of applications to a diverse set of applied sciences.

To evaluate the properties of the maximum likelihood procedure, one again turns to the pre-observation situation. Then $\widehat{\theta}(\boldsymbol{X})$ with the stochastic variable from the model inserted, is called the *maximum likelihood estimator*. For simplicity let us look at the situation with repeated independent continuous observations $\boldsymbol{X}_n=(X_1,...,X_n)$. Let $f_\theta(x)$ be the probability density of a single observation, and define the Fisher information by $I(\theta)=\mathrm{E}((\frac{\partial}{\partial\theta}\ln f_\theta(x))^2)$ assuming that this exists. Then under regularity conditions (see for instance Lehmann, 1999, where uniqueness of the local extrema is assumed for this), one can prove the following: 1) $\widehat{\theta}=\widehat{\theta}(\boldsymbol{X}_n)$ is consistent; 2) $\sqrt{n}(\widehat{\theta}-\theta_0)$ converges in law to $N(0,1/I(\theta_0))$ under θ_0, the true value, as $n\to\infty$. Thus the maximum likelihood estimator has some good asymptotic properties, and these results may be generalized to other cases with a large amount of data. However, there exist many examples of cases where the maximum likelihood estimator does *not* behave in an optimal way; see Le Cam (1990).

A good estimator $\delta(\boldsymbol{X})$ can also be found using a loss function $L(\delta(\boldsymbol{X}),\theta)$, for instance quadratic loss $L=(\delta(\boldsymbol{X})-\theta)^2$. One objective might be to minimize the risk, or expected loss, $R(\delta,\theta)=\mathrm{E}_\theta(L(\delta(\boldsymbol{X}),\theta))$, but a uniform minimization here is not feasible: Taking $\delta(\boldsymbol{X})\equiv\theta_0$ gives $R(\delta,\theta_0)=0$ for all reasonable loss functions. One way around this, is to limit oneself to unbiased estimators: $\mathrm{E}_\theta(\delta(\boldsymbol{X}))=\theta$ for all θ. The theory on this can be found in Lehmann and Casella (1998).

In addition to point estimation, statistical inference theory discusses hypothesis testing and confidence intervals. Hypothesis testing is closely related to confidence intervals. I will consider here one-sided confidence intervals $(-\infty,\bar{\theta}]$ and two-sided confidence intervals $[\underline{\theta},\bar{\theta}]$. The lower and upper limits of these intervals are functions of the data. When considered again in the pre-observational situation, they should have the properties

$$P_\theta(\theta\in(-\infty,\bar{\theta}(\boldsymbol{X})])=P_\theta(\bar{\theta}(\boldsymbol{X})\geq\theta)=\gamma, \tag{3}$$

$$P_\theta(\theta\in[\underline{\theta}(\boldsymbol{X}),\bar{\theta}(\boldsymbol{X})])=P_\theta(\underline{\theta}(\boldsymbol{X})\leq\theta\leq\bar{\theta}(\boldsymbol{X}))=\gamma, \tag{4}$$

where γ is a pre-assigned confidence coefficient, say 0.95 or 0.99.

The statistical methods discussed so far are called frequentist methods: They are coupled to a pre-observational distribution using the statistical model. The probabilities and expectations involved in this can be interpreted by a thought construction: Imagine that the whole experiment is repeated a large number of times. Then the imagined relative frequency of an event A for these repetitions is approximately equal to the hypothetical probability $P(A)$. The probabilities and expectations are therefore connected to the methods used and to the statistical model used.

There is another approach to statistical inference which has a long history, but has been particularly popular in the last few years: The Bayesian approach. Here the probabilities are imagined to be connected to the parameters themselves. The important assumption is that one first in some way has obtained a prior distribution on the parameter, say with a probability density $\pi(\theta)$. From this prior, one finds a posterior distribution, given the data, by using a variant of Bayes' formula

$$P(T|D) = \frac{P(T \cap D)}{P(D)} = \frac{P(T)P(D|T)}{P(D)}.$$

The first part of this formula is the definition of the conditional probability of T, given D. This definition is consistent with the Radon-Nikodym approach, and also consistent with what one calls conditional probability in simple examples. The second part of the formula is a consequence. Applied to a situation with a continuous parameter θ and a continuous data model with density $f_\theta(\boldsymbol{x})$, a formula for the posterior density of θ given the data is obtained:

$$\pi(\theta|\boldsymbol{x}) = \frac{\pi(\theta)f_\theta(\boldsymbol{x})}{f(\boldsymbol{x})} = \frac{\pi(\theta)f_\theta(\boldsymbol{x})}{\int \pi(\phi)f_\phi(\boldsymbol{x})d\phi}.$$

Consider first Bayesian point estimation. Again defining a loss function $L(\delta(\boldsymbol{x}), \theta)$, we can now introduce the Bayesian risk as

$$BR(\delta(\boldsymbol{x})) = \int L(\delta(\boldsymbol{x}), \theta)\pi(\theta|\boldsymbol{x})d\theta,$$

and find the estimate $\delta(\boldsymbol{x})$ which minimized BR. With quadratic loss this leads to the posterior mean $\int \theta\pi(\theta|\boldsymbol{x})d\theta$ as an estimate. Other possible estimates include the mode and the median of the posterior distribution, the mode being the maximum of the density and the median is the value such that the probability that the parameter is below this value, equals 1/2. In these estimates one can insert the pre-observational stochastic variable $\boldsymbol{X}$, compare them with estimators obtained by frequentists methods, evaluating estimators using a frequentist or Bayesian approach.

The Bayesian concept which replaces the confidence intervals is that of credibility intervals. Again consider the one-sided case $(-\infty, \theta^*(\boldsymbol{x})]$ and the two-sided case $[\theta_*(\boldsymbol{x}), \theta^*(\boldsymbol{x})]$. These intervals have direct interpretations in terms of

a probability distribution over the parameter, the posterior distribution:

$$P(\theta \in (-\infty, \theta^*(\boldsymbol{x})]) = \int_{-\infty}^{\theta^*(\boldsymbol{x})} \pi(\theta|\boldsymbol{x})d\theta.$$

$$P(\theta \in [\theta_*(\boldsymbol{x}), \theta^*(\boldsymbol{x})]) = \int_{\theta_*(\boldsymbol{x})}^{\theta^*(\boldsymbol{x})} \pi(\theta|\boldsymbol{x})d\theta.$$

To choose $\theta_*(\boldsymbol{x})$ and $\theta^*(\boldsymbol{x})$, this can again be given a preassigned value γ, say 0.95 or 0.99. In a specific sense, the interpretation of the credibility interval is simpler and more direct than the interpretation of the confidence interval.

There is much more to say about Bayesian theory and Bayesian methods; see Bernardo and Smith (1994), Box and Tiao (1973) and Congdon (2006).

The great weakness with the Bayesian approach is that the scientist should be able to specify a prior distribution of the unknown parameter. In a way he should be willing to and able to enter a wager on the values of this parameter. It is often claimed that if the scientist is not willing to do this, he should use an objective prior; for different formal ways to specify this concept, see Kass and Wasserman (1996). I have recently used such a prior myself (see Helland et al, 2011), but even so I would claim: There are many cases where the scientist could not or should not have any fully specified prior opinion about the parameter, even not one based upon symmetry or other 'objective' criteria. In such cases he should resort to frequentist methods. In statistical inference one should be flexible, not staying with one approach which should be imagined to cover all cases.

Quite recently there has been proposed a frequentist alternative to a distribution connected to a parameter: The confidence distribution; see Schweder and Hjort (2002) and Xie and Singh (2011). The idea is that one looks upon the confidence interval for any value of the confidence coefficient γ. Let $(-\infty, \tau(\gamma, \boldsymbol{x})]$ be a one-sided confidence interval with coefficient γ, where $\tau(\gamma) = \tau(\gamma, \boldsymbol{x})$ is an increasing function. Then $H(\cdot) = \tau^{-1}(\cdot)$ is the confidence distribution for θ. This H is a distribution function and has the property that $H(\tau(\gamma, \boldsymbol{X}))$ has a uniform distribution over the interval $[0, 1]$ under the model. According to Xie and Singh (2011), the distribution function H is to be looked upon as a distribution *for* the parameter, to be used in the epistemic process, not a distribution *of* the parameter, as we have in the Bayesian approach.

Three general book on statistical inference are Casella and Berger (1990), Bickel and Doksum (2001) and Cox (2006). For a discussion of Fisher's contributions with a view towards the future, see Efron (1998). The recent book by Cox and Donnelly (2011) discusses many aspects of applied statistics and also provides some links to theoretical statistics.

2.4 Inference for discrete e-variables.

Opinion polls, or sample surveys, while very much used in practice, are not much discussed in the standard mathematical statistical literature. But specialized

books like Cochran (1977) exist. The framework is that one has a finite population consisting of N units, say N human beings, and one seeks some information about this population from a smaller sample of n units. The simplest case is when the sample is taken randomly from a register of the whole population, but other sampling plans exist. To increase efficiency, one often uses stratified sampling: The population is divided into k strata using some relevant criterion, such that the numbers of unit in stratum i is N_i, and one samples randomly n_i units from this stratum. Of course $\sum_{i=1}^{k} N_i = N$ and $\sum_{i=1}^{k} n_i = n$.

As a simple epistemic problem, assume that an unknown number M in the population possesses some specific property A, and one wants to use the sample to estimate $\theta = M/N$. This θ takes a discrete set of values $0, 1/N, ..., N/N$, and is not always called a parameter. In this book we will use the more general concept of an e-variable, a conceptual variable which is unknown before the epistemic process begins. In general it is implicit in the concept of an e-variable that this is a quantity that we want to gain knowledge about.

A more general problem is that each unit j in the population has some value y_j attached to it, and one wants to estimate $\theta = \bar{y}_{population} = \sum_{j=1}^{N} y_j/N$. The simple problem above is then obtained by specializing y_j to be an indicator function. A common estimate of θ is

$$\widehat{\theta} = \frac{\sum_{sample} y_j/\pi_j}{\sum_{sample} 1/\pi_j},$$

where π_j is the probability that unit j should be included in the sample. This can be used for many sampling plans. For stratified sampling we get $\widehat{\theta} = \sum_{i=1}^{k} N_i \bar{y}_i/N$, where the mean $\bar{y}_i$ is over the sample in stratum i.

Opinion polls are based on the assumption that each person 'has' an opinion on the issue that is focused upon. The fact that opinions may vary with time, and that they may depend on the contexts, is perhaps realized, but it is not much discussed in this connection. To see this first from the point of view of the person being interviewed, imagine for instance that a woman A has spent some time on an hotel, and then after a few days receives a questionnaire by e-mail, one of the questions being: 'On a scale from 1 to 10, how do you evaluate the service at this hotel?' This causes her to enter an epistemic process, mostly related to introspection. To begin with, the score is some unknown number θ, but after a while she decides on a value $\theta = u_k$, one of the values from 1 to 10.

This decision process may be evaluated subjectively by the woman A herself. We may also consider the whole situation as looked upon by some person B knowing her background. In this latter case prediction is relevant. The person B may have access to some kind of data from a sample of size n_i from a stratum consisting of people with the same background as the woman, and to detailed information about the hotel. On this basis he may want to predict θ. Again this is an epistemic process with a discrete e-variable; the target for the prediction of this e-variable is not a population, but a single unit, the woman A in this case. B may wish a large n_i to have accurate data, but at the same time resources may be limited: He may be forced to have a small n_i in order to be able to predict from a fairly homogeneous subpopulation.

Both in the introspection case and in the external observer case, one can consider the following: Assume that the woman A had an unfortunate episode with the receptionist of the hotel just before she left, and that the observer B does not know about this episode. Let θ' be the hypothetical score that A would have had if the episode had not taken place. Depending on the further circumstances, θ' may not be accessible to the woman A herself, and θ may not be accessible to the observer B.

Considerations of this kind are here very vague, but they should give some feeling of what I mean by an epistemic process when a single unit is involved. The situation here is on the borderline or outside what one meets in ordinary statistics, but the point is that it describes epistemic processes, and that one of these processes (the prediction part) in principle can be made precise in statistical terms. Considerations of this kind will be important when I later will try to approach the foundation of quantum mechanics from an intuitive point of view.

The Bayesian concepts of prior and posterior distribution are straightforward to formulate in the case of a discrete e-variable, and the concept of confidence distribution also carries over: If θ takes the values $u_1, ..., u_r$, then the confidence coefficient γ can take only r values, and the confidence distribution H is determined as follows: Let again $(-\infty, \tau(\gamma, \boldsymbol{x})]$ be a one-sided confidence interval with coefficient γ, where $\tau(\gamma) = \tau(\gamma, \boldsymbol{x})$ on the r values. Then $H(\cdot) = \tau^{-1}(\cdot)$ can be extended to a discrete distribution function for θ, which has the property that $H(\tau(\gamma, \boldsymbol{X}))$ for data $\boldsymbol{X}$ has a uniform distribution on the r values $H(u_1), H(u_2), ..., H(u_r) = 1$.

3 Group actions and model reduction.

In simple random sampling, a natural objective prior for the e-variable θ is found by giving the same probability $1/N$ to each unit in the population. This is the invariant measure (see Appendix 3) for the permutation group. In general an epistemic problem related to a θ may often have some symmetry property associated with it, and this is formalized by introducing a group of transformations acting upon the space Θ of the e-variable. When θ is transformed by the group and the observations are transformed accordingly (see Helland, 2004), one should get equivalent results from the statistical analysis. As a trivial example: One should get equivalent results from a statistical analysis whether the parameters and the observations are measured in meters or in centimeters. A summary of basic group theory is given in Appendix 3. Examples of groups acting upon a parameter space, are location: $\xi \to \xi + a$ for a real; scale group: $\sigma \to b\sigma$ for $b > 0$; location and scale: $(\xi, \sigma) \to (a + b\xi, b\sigma)$; rotation in a multidimensional parameter space; a general linear group acting upon a multidimensional parameter space etc.. Invariance under a group may help improving the estimation or the inference in general.

When we later come to our link to quantum mechanics, there seems in that case to be a canonical choice of the group G. At present, we just say vaguely

that we choose G, if possible, in agreement with some symmetry aspect of the whole situation.

Now fix a point θ_0 in the e-variable space Θ. An *orbit* in this space under G is the set of points of the form $g\theta_0$ as g varies over the group G. The different orbits are disjoint, and θ_0 can be replaced by any e-variable on the orbit. Any set in Θ which is an orbit of G or can be written as a union of orbits, is an invariant set under G in Θ, and conversely, all invariant sets can be written in this way. If there is only one orbit in Θ, the group is said to be acting transitively upon Θ.

A statistical model should be as simple as possible, but not simpler. In some cases we may want to do a simplification, a model reduction. This may take the form of a reduction of the e-variable space Θ. Parts of this space which are essential for the epistemic process, must always be retained, but irrelevant dimensions should be left out. I will now formulate a general criterion which will be used throughout this book:

If there is a group G acting upon the e-variable space Θ, any model reduction should be to an orbit or to a set of orbits of G.

This will ensure that G also can be seen as a group acting upon the new e-variable space. In particular, if the group actions form a transitive group G, no model reduction is possible.

EXAMPLE 1. Assume that a single set of observations is modeled by some large parametric model, only assuming that parametric class contains the normal model. Let the location and scale group be acting upon the parameter space Θ. Then one orbit is given by the $N(\xi, \sigma^2)$ distribution. This is not an uncommon model reduction.

EXAMPLE 2. Look at two independent sets of observations: $(X_1, ..., X_m)$ independent and identically $N(\xi_1, \sigma_1^2)$ and $(Y_1, ..., Y_n)$ independent and identically $N(\xi_2, \sigma_2^2)$. Let G be the translation and scale group given by $\xi_1 \to a_1 + b\xi_1,\ \sigma_1 \to b\sigma_1,\ \xi_2 \to a_2 + b\xi_2,\ \sigma_2 \to b\sigma_2$. Note that a common scale transformation by b is assumed. Then the orbits of the group in the parameter space are given by $\sigma_1/\sigma_2 = \text{constant}$. A common model reduction is given by $\sigma_1 = \sigma_2$. This simplifies the comparison of ξ_1 and ξ_2, which is often the goal of the investigation.

EXAMPLE 3. Linear statistical models have a large range of applications. In general these models have the form where the observations Y_l are independent $N(\xi_l, \sigma^2)$, where the expectations ξ_l are linear combination of a set of parameters. In one particular such model (the two-way analysis of variance model) the observations Y_{ijh} have expectations $\mu + \alpha_i + \beta_j + \gamma_{ij}$. To get a unique representation of this kind, one often imposes the restrictions $\sum_i \alpha_i = 0$, $\sum_j \beta_j = 0$, $\sum_i \gamma_{ij} = 0$ for each j and $\sum_j \gamma_{ij} = 0$ for each i. Let the group G be given by all permutations of the index i and all permutations of the index j. Then

an obvious model reduction is given by the invariant set where the expectation is $\mu + \alpha_i + \beta_j$. This is called the model without interaction, and is a valid simplification in some cases.

EXAMPLE 4. Another example of a linear model is the polynomial regression model $Y_i = \beta_0 + \beta_1 x_i + ... + \beta_p x_i^p + E_i$, where the E_i's are independent $N(0, \sigma^2)$ for $i = 1, ..., n$. Let G be the group defined by translations in the x-space: $x \to x + a$, which generates a transformation group on the parameters $(\beta_0, ..., \beta_p)$. Then the submodels $Y_i = \beta_0 + \beta_1 x_i + ... + \beta_q x_i^q + E_i$ $q < p$ correspond to invariant sets in the parameter space.

EXAMPLE 5. A further example of a linear model is the multiple regression model $Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + E_i$ for $i = 1, .., n$ with fixed x_{ij}, which again has many different applications. Consider first the case where the x_{ij} are measured in different units for different j. Then there is a natural transformation group given by separate scale changes $x_{ij} \to k_j x_{ij}$ $(j = 1, ..., p)$. This induces a group on the regression parameters by $\beta_j \to \beta_j / k_j$ $(j = 1, ..., p)$. The invariant sets in the parameter space are found by putting some of the β_j's equal to 0. These reduced models are well-known from many applications of regression analysis.

EXAMPLE 6. Consider the same multiple regression model as in Example 5, but assume now that the explanatory variables x_{ij} all are measured in the same units. A large class of transformations $\boldsymbol{x}_{i\cdot} \to \boldsymbol{Q}\boldsymbol{x}_{i\cdot}$ may then be of interest. In particular, an interesting case is when $\boldsymbol{Q}$ varies over the orthogonal matrices.

As here, and as in any linear model, estimates of the regression parameters can in principle be found by the method of least squares, which is equivalent to the maximum likelihood method. However, this method breaks down when $p > n$, or more generally when one has collinearity problems such that the matrix which we need to invert in order to implement the least squares solution, is singular. A large number of alternative estimation methods are proposed in the statistical literature to tackle this problem, but it seems very difficult to decide which of these methods one should use in practice.

For this problem, one place where one may start the investigation, is that many of the methods are *equivariant* under the transformation induced by rotation in the x-space: A transformation on $\widehat{\theta}$ found from transformations of the data is the same the corresponding transformation on the parameter θ.

Before I return to this problem, I will summarize a little more theory. In Appendix 3 the concept of a right invariant measure for the group is defined, and it is recommended that such a prior is used as an objective prior. Among other things it is proved in Helland (2004, 2010) that there for a transitive group is a very close connection between confidence intervals and Bayesian credibility intervals in this case. It follows from this that there is a close connection between confidence distributions and posterior distributions with this prior.

Concerning equivariant estimators, there is a generalization of an old theorem by Pitman, which is proved in Helland (2010), showing that if the loss function is invariant and proportional to the quadratic loss, if the group is transitive and a right invariant prior is used, then the posterior mean, if finite, is the best equivariant estimator.

EXAMPLE 6, CONTINUED. Look at a modification of the model in Example 6 where the explanatory variables are random variables X_{ij}. This is natural in many observational studies. For simplicity, assume that all variables are centered: $\mathrm{E}(Y_i) = 0$ and $\mathrm{E}(X_{ij}) = 0$. Then the model is $Y_i = \beta_1 X_{i1} + ... + \beta_p X_{ip} + E_i$ for $i = 1, .., n$. Let $\boldsymbol{\Sigma}_x$ be the covariance matrix of the x-variables, which can be defined by the property that $\mathrm{Var}(\sum_j a_j X_{ij}) = \boldsymbol{a}^T \boldsymbol{\Sigma}_x \boldsymbol{a}$ for all vectors $\boldsymbol{a} = (a_1, ..., a_p)^T$. Then $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ can always be expanded in terms of an orthogonal set of eigenvectors $\boldsymbol{d}_i$ of $\boldsymbol{\Sigma}_x$:

$$\boldsymbol{\beta} = \sum_{j=1}^{p} \gamma_i \boldsymbol{d}_i. \tag{5}$$

In this expansion the number of terms can be reduced in two ways: 1) Some of the eigenvalues may be coinciding. Then the eigenvectors in this eigenspace can be rotated in such a way that there is just one eigenvector in this space which has a nonzero component along $\boldsymbol{\beta}$. 2) The vector $\boldsymbol{\beta}$ has no component in this eigenspace. So an interesting reduced model is the one with m non-zero terms in (5). The ordering of the terms in (5) is arbitrary, so the reduced models only specify the number m of non-zero terms, not which terms that are non-zero. It is not difficult to show that these models for different m are exactly the orbits of the following group G: Rotations in the x-space and hence of the eigenvectors $\boldsymbol{d}_i$ augmented by independent scale transformations $\gamma_i \to a_i \gamma_i$ where $a_i > 0$.

It is shown in Helland et al. (2012) and Cook et al. (2012) that these reduced models coincide with reduced models introduced by researchers from two different traditions: The envelope model of Cook et al. (2010) and a natural population model arising from the partial least squares algorithmic 'soft' models from chemometrics. Maximum likelihood estimation and other estimators under the reduced model are discussed in Cook et al. (2012) and Bayes estimation in Helland et al. (2012). The invariant prior induced by the group leads to an undefined posterior expectation, so a best equivariant estimator can not be found from this. However approximating the scale prior with a proper prior leads to $\boldsymbol{\beta}$-estimates, hence predictions, which seem to have good properties.

Finally I give for completeness a simple example of model reduction for the case where the e-variable in question is not a continuous parameter.

EXAMPLE 7. In stratified random sampling, the natural group G is the group of independent permutations within each stratum. The orbits of G are then given by the single strata, and model reductions to invariant sets are given by reduction to any set of strata. Such a reduction is of course natural in cases where one want to limit the investigation to a particular set of strata.

4 Interlude

Before the epistemic process, one has an unknown e-variable θ. What is the situation after one has gone through an epistemic process? In the case where θ is the parameter of a statistical model, the situation can be summarized as follows: Depending upon the statistical philosophy used, one has either a confidence interval $[\underline{\theta}(\boldsymbol{x}), \bar{\theta}(\boldsymbol{x})]$ or a credibility interval $[\theta_*(\boldsymbol{x}), \theta^*(\boldsymbol{x})]$. In both cases, assume that the coefficient γ is very high, say 0.999. Then the practical conclusion from the epistemic process is that the new state is specified by saying that θ belongs to this interval.

Confidence intervals or credibility intervals for discrete e-variables are not much discussed in the statistical literature, but the concepts carry over. One difference is that when one has very much data, the intervals can degenerate into a single point. In the following, it will make the discussion much simpler to consider such a case. Assume that one has this situation, and let again the coefficient γ is very high, say larger than 0.999. Then in the frequentist case, one has a conclusion of the type $P_\theta[\widehat{\theta}(\boldsymbol{X}) = u_k] = \gamma$ with realized data $\boldsymbol{X} = \boldsymbol{x}$, and in the Bayesian case one has a posterior probability $P[\theta = u_k|\boldsymbol{x}] = \gamma$. In both cases we conclude for practical purposes that the new state is given by $\theta = u_k$, and that this value can be used in further investigations.

Any epistemic process starts with an unknown e-variable θ, and when the process ends, one has some knowledge about θ. A state is obtained when this knowledge is almost certain. In the simplest case the knowledge can be expressed by a certain fixed value $\theta = u_k$. This situation can be realized by a statistical investigation with a discrete parameter/e-variable, but it can also be realized in other epistemic situations. One example is when a person through introspection makes up his or her mind on a particular issue, as illustrated with the woman answering an opinion poll in subsection 2.4. Again we can talk about a state when the person's knowledge about his/her opinion is almost certain. Looking upon the process of achieving an opinion on an issue as an epistemic process can of course be discussed; in any case it involves philosophical and psychological questions that are beyond the scope of the present book. However, many examples of everyday epistemic processes can be given, some realized through the communication with other people. Some of these processes start with an unknown θ and end up with an (almost) sure state $\theta = u_k$. Other examples of this are connected to prediction of some variable. In these last examples, the e-variable is typically attached to a single unit, not to a population of units, which is the most common situation in statistical investigations.

The example in Section 2.4 illustrates another issue: Here one e-variable θ is accessible to the woman A herself, while another e-variable θ' (the hypothetical score if a certain episode had not taken place) is accessible to the person B knowing her background and having information about the hotel. The reason for this difference is that the two persons have different background knowledge. We will come back to similar situations later when discussing quantum mechanics.

Returning now to statistics, nearly all papers on statistical inference have data models, that is, either parametric or nonparametric models of the observed

data, as their point of departure. Also, statistical practice is deeply founded upon this tradition. Even though a different culture was promoted and discussed by Breiman (2001), the data modeling culture is now more dominant than ever.

Nor will this paper depart radically from this culture, but we will add an element to it: Every decision in any experimental or observational setting is made in a context. This context may not be trivial, and may have decisive influence on how the inference should be made. The context may be in parts be formed by the historical and cultural background for the study, and it may depend upon earlier decisions. But it can also in addition be conceptual, including the formulated goal of the investigation, the model, a loss function and/or a Bayesian prior. Also the framework for the study must be considered as a part of the context: The experimental units available, what can be measured on these units; limitations in terms of money, time and human resources.

In order to be able to discuss contexts in general, it turns out to be useful first to give a precise definition of what we mean by conceptual variables, which includes observations, parameters, latent variables and more. Then we define e-variables, which up to now has been a loosely defined concept.

In statistical theory, a parameter is often defined as an index of a class of distributions, but in statistical practice, a parameter is often a quantity of interest in itself, introduced as an expectation, a variance, a covariance, a correlation, a regression coefficient or a probability. These two facets of the parameter concepts may to some extent be regarded as complementary, even though this introduces no logical difficulty.

In the statistical tradition, a parameter is usually connected to a hypothetical infinite population, but in fact parts of the statistical theory - in reality everything that is not related to asymptotical considerations - can be generalized to the case where the parameter is replaced by a conceptual variabel connected to a single unit or to a few units.

In addition to the unknown conceptual variable there are data $\boldsymbol{x}$. The purpose of an experiment is then to use these data to answer questions formulated in terms of the conceptual variable. This will be the background for our approach to essential parts of quantum theory later in this book.

5 Conceptual variables and contexts

Fisher (1922) introduced the concept of parametric models in the way it is used throughout statistics today. According to Stigler (1976) and Cook (2007), the word 'parameter' is mentioned 57 times in that groundbreaking paper. Recently, Taraldsen and Lindqvist (2010) argued that in Bayesian inference the parameters and the potential observations should be defined on the same underlying probability space. This is one point of departure of the present book. Another point of departure is that in any situation where inference is supposed to be done, several other types of unknown variables than parameters are of relevance (one simple example is in prediction), and that additional types of variables are needed to describe the context of the experiment or observational study.

DEFINITION 1. *Consider any experimental or observational situation at a given time or over some time span, more generally any epistemic process. Any variable which can be formulated conceptually by a person or by a group of persons in that situation is called a conceptual variable.*

This term was indicated in Helland (2010), where it also was argued that some unknown conceptual variables could be inaccessible, that is, they could not be assessed with arbitrary accuracy through estimation or prediction in any way in the given situation. This was taken as the first steps in a line of reasoning indicating a connection between theoretical statistics and quantum theory, a line of reasoning that I will continue below.

In the following, I may alternately speak about one conceptual variable and several related conceptual variables in the same way as we may talk about a multivariate parameter or several one-dimensional parameters.

Several classes of conceptual variables are of interest in a statistical investigation, depending upon the situation:

- Context variables: The background variables for an experiment or observational study.
- OCV's (observed conceptual variables): Data or preset values.
- Statistics: Known functions of the data.
- Quasi-statistics: Known or unknown functions of the data.
- Input variables and responses/output variables. As used in prediction and regression, cp. Hastie, Tibshirani and Friedman (2009).
- UCV's (unknown conceptual variables): For instance parameters, latent variables or a response for a new set of input variables.
- Hypothesis variables: Concepts from which one may formulate assertions about the value of a parameter.
- Conclusion variables: Conceptual variables by which one may formulate the conclusions from an experiment or observational study.

In this book I will consider any epistemic process.

DEFINITION 2. *A conceptual variable which is used in an epistemic process is called an* e-variable θ. *Before the epistemic process is started, the e-variable is unknown. After the process, one is able to achieve some conclusion about the e-variable, the simplest case being that we know its value:* $\theta = u_k$, *a type of conclusion which is only possible for discrete e-variables.*

We also have the important concept of a *context* of an epistemic process.

DEFINITION 3. *In the case of an experiment the context includes the setting of the experiment, similarly for an observational study. But in general for any epistemic process it also includes the background for the process, historical, specific and conceptual. The conceptual background for any study should always include a formulated goal of the study.*

In a series of experiments or in a meta-analysis, the conclusions from one situation may be used as a part of the context of the next situation.

Several operations may be done on any assertion containing conceptual variables, including $\neg$ (negation), $\wedge$ (and) and $\vee$ (or). Formulating statements connected to a concrete experimental or observational situation may then be done using propositional logic, a subject which has a large abstract literature. As formulated in Appendix 6, I want to be more concrete and regard sentences formulated in ordinary, everyday language as primitive entities.

There is a close connection between propositional logic and set theory, where we identify $\neg$ with complement, $\wedge$ with intersection and $\vee$ with union. Such identifications are often done implicitly in elementary textbooks in probability. Let $(\Omega', \mathcal{F}')$ be the measurable space thus obtained, where $\mathcal{F}'$ is a σ-algebra of subsets of Ω'. On some measurable subset Ω of Ω' one can define conditional probability measures related to one conceptual variable given other conceptual variables, where the conceptual variable conditioned upon may or may not belong to Ω, that is, may or may not be measurable functions on $(\Omega, \mathcal{F})$, where $\mathcal{F} = \{A \cap \Omega : A \in \mathcal{F}'\}$. Strictly speaking, conditioning here must be taken as more general than the usual conditioning in statistics where we condition upon σ-algebras. We are talking in general about probabilities, given some *information*, so that we should wish to stay within the framework of propositional logic. As indicated in Appendix 6, however, it seems like we need some extra assumptions in this framework to make the conditional probabilities precise in general. Therefore I will in this book stay within the probabilistic framework and limit myself to conditional probabilities given a σ-algebra as defined by (1). Conditional probabilities, given some conceptual variable τ which is a random variable on $(\Omega, \mathcal{F})$, is defined as the conditional probability, given the σ-algebra generated by this conceptual variable, that is, the collection of sets $\tau^{-1}(B)$, where B runs through the relevant Borel sets. Conditional probabilities, given some non-random variable τ are simply measurable functions of this variable.

When considering conditional probabilities, given the context, in most cases only part of the context will be relevant. The conceptual variables on which probabilities can be defined, will be called random variables. For simplicity, technical problems resulting from the fact that conditional distributions are only defined almost surely, are mostly disregarded in this book. However, difficulties from this in the definition of sufficiency (see Lehmann and Casella, 1998) will be addressed.

A *statistical model* is defined as a conditional distribution of the data, given all parameters (together with the context, including preset values). It is assumed as usual that this class is dominated, that is, all conditional distributions are absolute continuous with respect to some fixed conditional probability measure

P, given the context, where Q is defined to be absolutely continuous with respect to P if $P(A) = 0$ implies $Q(A) = 0$.

In addition, if a Bayesian analysis is to be carried out, there is a prior distribution of of the parameters (again given the context). To allow for objective priors, I will, in agreement with Taraldsen and Lindqvist (2010) allow these measures to be unnormalized; see that paper and also the recent paper by McCullagh and Han (2011) on how logical difficulties with this can be avoided. Note that I talk about a Bayesian analysis to be carried out, not about Bayesian or frequentist research workers. The same person may in certain cases carry out both types of analysis, first a frequentist analysis and then at a later point of time a Bayesian analysis.

In the following, I will depart from my earlier notation and also denote random data by lower case letters. It will be clear from the context whether I talk about the pre-experimental or post-experimental situation. The statistical model will, if this is natural, be seen from a pre-experimental point of view.

In the first part of this book I will in particular address the following prediction or learning situation: In the statistical model, let y_i have some identical conditional distribution, given x_i and some fixed parameter θ for $i = 0, 1, 2, ..., n$, and assume that these distributions are independent. In addition x_i $(i = 0, 1, 2, ..., n)$ may or may not have some identical independent distributions given a parameter κ, and θ and κ may or may not have priors. I assume that y_0 is unknown, but the other y_i's are observed data. The x_i's are data or preset values. Thus here the UCV's are y_0, θ and κ, while the OCV's are $\{x_i, y_i; i = 1, ..., n\}$ and x_0. In principle the variables may belong to any topological space and the σ-algebras of relevance may be contained in the Borel σ-algebra, but in most practical cases they are constrained to subsets of Euclidean spaces. My goal in this part is to give some theoretical basis for discussing methods to predict y_0, given the OCV's. Thus y_0 is in this case the e-variable of interest. This is also the conceptual basis for much of Hastie, Tibshirani and Friedman (2009) (supervised learning).

6 Data; generalized sufficiency and ancillarity

Let z be a statistic, and let τ be a conceptual variable. The following is assumed throughout this section:

1) The distribution of z, given τ, depends on an unknown e-variable θ.

2) If τ or part of τ has a distribution, this is independent of θ. The part of τ which does not have a distribution is functionally independent of θ.

To fix ideas, think of 1) and 2) as describing a situation where inference on θ is sought from the data z in the context described by τ, but there are variants of this. In a simple experiment, z may be the whole data set, and τ may be trivial or some nuicance parameter. In addition, τ may contain the real context of the experiment, which it always will, but this is often just taken as

an implicit fact. In a series of experiments, ordered in time, z may be the data set of the last experiment, and the context τ may contain some or all of the conceptual variables connected to the earlier experiments. In a metaanalysis, z may contain all data sets, and τ may contain all contexts. It is a basic condition that the model assumptions are rich enough so that 1) and 2) are meaningful. Throughout most of this section, z and τ will be held fixed.

6.1 Sufficiency

We let t be a known or unknown function of z. Later I will give a class of examples of the perhaps unfamiliar situation where we have an unknown function of the data. The concept of sufficiency was introduced by Fisher as a tool for reducing the data in a given situation without sacrificing anything related to the inference on the parameter θ.

DEFINITION 4. *We say that $t = t(z)$ is a (z, τ)-sufficient quasi-statistic for θ if the conditional distribution of z, given t, τ and θ is independent of θ. If z is the whole data set, we say just that t is τ-sufficient.*

From the fact that 1) is meaningful, it follows that the conditional distribution of z, given t, τ and θ is meaningful. However, difficulties (Lehmann and Casella, 1998) may arise because the conditional distribution is only defined almost everywhere. I then follow Reid (1995) in making the definition more precise: The quasistatistic $t(z)$ is (z, τ)-sufficient if there is a transformation from z to (t, v) such that the densities satisfy

$$f(z|\theta, \tau) \propto f(t|\theta, \tau) f(v|t, \tau),$$

where the constant of proportionality is independent of θ. This is a version of the factorization theorem: $t(z)$ is (z, τ)-sufficient if and only if there exist functions $g(t|\theta, \tau)$ and $h(z|\tau)$ such that for all z and θ we have

$$f(z|\theta, \tau) = g(t(z)|\theta, \tau) h(z|\tau).$$

Ordinary sufficiency results if τ is trivial, θ is the full parameter and t is a statistic. The case where part of τ is a nuicance parameter is also of interest. The general concept is of interest also in many other situations.

In general, if $t(z)$ is a (z, τ)-sufficient statistic, the rest of the distribution of z can be thought of as generated by some randomization independent of θ, and gives no information about the e-variable. This will be made precise by a sufficiency principle formulated later.

It is clear that $t = z$ is a (z, τ)-sufficient statistic, but usually we are interested in smaller functions of z. In general a minimal sufficient observator will not exist, but translating a result from ordinary sufficiency theory, any boundedly complete (z, τ) observator will be minimal sufficient.

DEFINITION 5. *A (z,τ)-sufficient statistic t is boundedly complete if for all bounded functions h*

$$E(h(t)|\theta,\tau)=0 \text{ for all } \theta \text{ implies } P(h(t)=0|\theta,\tau)=1 \text{ for all } \theta.$$

PROPOSITION 1. (Bahadur's Theorem). *Suppose that t takes values in a k-dimensional Euclidean space and that t is a (z,τ)-sufficient and boundedly complete statistic. Then t is a minimal (z,τ)-sufficient statistic.*

Standard results like the Rao-Blackwell Theorem and the Lehmann-Scheffé Theorem generalize immediately to (z,τ)-sufficiency. The first result says that if $g(z)$ is any estimator of θ and if $t(z)$ is a (z,τ)-sufficient statistic, then the conditional expectation of $g(z)$, given $t(z)$ is an at least as good estimator as $g(z)$, using quadratic loss. Sometimes one gets a considerable improvement using such a Rao-Blackwellization. The last result says that if t is complete and τ-sufficient for θ and $h(t)$ is an estimator of θ which is conditionally unbiased, given τ, then $h(t)$ has uniform minimal conditional variance, given τ.

Assume that we on the basis of data z want to estimate the e-variable θ in the context given by τ.

DEFINITION 6. *If $t(z)$ is a minimal sufficient quasi-statistic for θ, and the distribution of t depends on a part of τ , we say that this part is relevant for the estimation of θ.*

EXAMPLE 8. Let $z=(y_1,...,y_n)$, where $y_1,...,y_n$ are independent and identically distributed (i.i.d.) $N(\mu,\sigma^2)$. Then $(\bar{y},s^2)$ is sufficient for (μ,σ^2), where $s^2=(n-1)^{-1}\sum_{i=1}^n(y_i-\bar{y})^2$. However, even in this simple example it is of interest which parameter we focus upon. Write the log likelihood as

$$\ln f = k+\frac{1}{2\sigma^2}\sum_{i=1}^n(y_i-\mu)^2+\frac{1}{2}\ln(\sigma^2) = k+\frac{1}{2\sigma^2}\sum_{i=1}^n[(y_i-\bar{y})^2+n(\bar{y}-\mu)^2]+\frac{1}{2}\ln(\sigma^2).$$

From this we see:

a) If τ contains the nuisance parameter σ^2, then $\bar{y}$ is (minimal) (z,τ)-sufficient for μ.

b) If τ contains the nuisance parameter μ, then $\sum_{i=1}^n(y_i-\mu)^2$ is (minimal) (z,τ)-sufficient for σ^2. This is a first example of an unknown function of the data where the concept of sufficiency is of interest.

In each case the minimality of the sufficient statistic can be proved from Proposition 1.

We conclude from this that σ^2 is irrelevant for the (point-)estimation of μ, while μ is relevant for the estimation of σ^2.

6.2 Ancillarity and conditioning

Next I turn to the generalization of ancillarity, another basic concept introduced by Fisher.

DEFINITION 7. *We say that $u = u(z)$ is a (z, τ)-ancillary quasi-statistic for θ if the conditional distribution of u, given τ is independent of θ.*

If u is (z, τ)-ancillary and f is a measurable function, then $f(u)$ is (z, τ)-ancillary. In the corresponding partial ordering of statistics ($u < v$ if $u = f(v)$ for some function f), z is an upper bound. By Zorn's Lemma, one or several maximal ancillaries will exist. We say that u is τ-ancillary if z is the whole data set; just ancillary if τ is trivial.

A very important and much discussed question is when one should condition upon ancillaries. Once one has conditioned upon an ancillary, this can be taken as part of the context of the experiment or the observational study. Thus the context is expanded, but after this expansion, 1) and 2) in the beginning of Section 6 will still hold. A closer discussion of the question of conditioning will be given in the next section, but as a background for this discussion we will sketch some examples.

A basic argument for conditioning is given by the following example:

EXAMPLE 9. (Berger and Wolpert, 1988; Cox, 1958) Consider two potential laboratory experiments for the same unknown parameter θ such that $\mathcal{E}^1$ is planned to be carried out in New York while $\mathcal{E}^2$ is planned to be carried out in San Francisco. The owner of the material to be sent chooses to toss an unbiased coin, deciding $\mathcal{E}^1$ with probability 1/2 and $\mathcal{E}^2$ with probability 1/2. Consider the whole experiment $\mathcal{E}$ including the coin toss, and let u be the result of the coin toss. Here u is ancillary, and everybody would condition upon u in the statistical analysis.

A problem with the requirement of conditioning, is that maximal ancillaries may not be unique.

EXAMPLE 10. Let θ be a scalar parameter between -1 and +1. Consider a multinomial distribution on four cells with respective probabilities $p_1 = (1 + \theta)/6$, $p_2 = (2 - \theta)/6$, $p_3 = (1 - \theta)/6$ and $p_4 = (2 + \theta)/6$ and total number of observations n. Let the corresponding observed numbers in the sample be z_1, z_2, z_3 and z_4. The multinomial distribution is a generalization of the binomial distribution with multivariate point probabilities

$$\frac{n!}{z_1!z_2!z_3!z_4!} p_1^{z_1} p_2^{z_2} p_3^{z_3} p_4^{z_4}.$$

Then one can show that each of the statistics

$$u_1 = z_1 + z_2, \quad u_2 = z_1 + z_3$$

is ancillary for θ, but they are not jointly ancillary. And conditioning upon u_1, respectively u_2 leads to distinct inference (the maximum likelihood estimator is the same, but the asymptotic variances are different).

Cox (1971) has proposed an intrinsic criterium for the choice of ancillary to condition upon in such cases, but my opinion is that this choice should depend upon the context.

EXAMPLE 11. In a certain city the sex ratio is 1:1, and it is known that 1/3 of the population have their own cellphone. The ratio between female and male cellphone owners is an unknown quantity $(1+\theta)/(1-\theta)$, where $-1 < \theta < 1$. One is interested in estimating θ by sampling randomly n persons from a register of the city population. It is assumed that the population is much larger than the sample size n.

Let the number of men in the sample be u_1, and let u_2 persons in the sample be owners of cellphones. Thus $u_1 = z_1 + z_2$ and $u_2 = z_1 + z_3$, where z_1 and z_2 are the male cellphone owners and non-owners, respectively, and z_3 and z_4 are the corresponding female numbers. The joint distribution of $z_1, ..., z_4$ is as in Example 10. Again each of u_1 and u_2 are ancillary, but they are not jointly ancillary.

Another question is whether or not one should *always* condition upon ancillaries. The following examples give a background for that discussion.

EXAMPLE 12. (Helland, 1995) As a part of a larger medical experiment, two independent individuals (1 and 2) have been on a certain diet for some time, and by taking samples at the beginning and at the end of that period some response like the change in blood cholesterol levels is measured. For the individual i $(i = 1, 2)$, the measured response is y_i, which is modeled as independent normal (μ_i, σ^2) with a known measurement variance σ^2.

Because the two individual have been given the same treatment (diet) in the larger experiment, the parameter of interest is not μ_1 and μ_2, but their mean: $\theta = \frac{1}{2}(\mu_1 + \mu_2)$.

Suppose now that for some reason we have only capacity to measure one of the individuals, but at the outset, we don't know which. Let u be the indicator of the individual chosen. It is clear that, given u, that is, given $u = 1$ or $u = 2$, we get no information about θ. But choosing u randomly with probability $\frac{1}{2}$ for each of the two values will give us such information, provided that the identity of the individual chosen is not revealed. The last statement follows from a sampling argument: The situation is a special case of a sampling situation where n individuals are sampled randomly from a population of N individuals where the parameter of interest is the mean in the population. In addition there is a measurement error for each individual, modeled as independently normal $(0, \sigma^2)$. It is then clear that the sample mean of the observations is an appropriate estimator of this population mean. It is equally clear that this

conclusion also must be valid for the special case $N = 2, n = 1$, the situation at hand.

The surprising aspect of this example is that a situation with less information can give us more ability to do inference: By not knowing u we can make some (admittedly uncertain, but nevertheless valid) inference on θ; when we know u, such an inference is impossible.

EXAMPLE 13. Consider a sensory analysis firm where there is a staff of N trained assessors and a panel of n out of these are selected randomly to taste a particular product. A report is written. Given the assessors, what they do in the analysis must be considered as separate experiments on a common parameter θ. Consider the whole investigation, and let u be the result of choosing randomly the assessors to take part in it. Again u is ancillary. But in this case it may not be immediately natural to condition upon u in the written report.

EXAMPLE 14. (This example requires knowledge of some special statistical procedures.) Look at the general comparison of logistic regression and linear disciminant analysis (Hastie et al., 2009 and Efron, 1975). With inputs $\boldsymbol{x}$ and classification into one of K possible classes g, both procedures correspond to the same linear form $\beta_{k0} + \boldsymbol{\beta}_k^T \boldsymbol{x}$ of

$$\log \frac{P(g = k|\boldsymbol{x})}{P(g = K|\boldsymbol{x})}.$$

The difference is that in logistic regression, β_{k0} and $\boldsymbol{\beta}_k$ are the only parameters, $\boldsymbol{x}$ is ancillary, and inference is done conditionally upon $\boldsymbol{x}$ in the training set; while in linear discriminant analysis (LDA), these parameters depend upon further parameters characterizing the underlying assumed multinormal distribution of $\boldsymbol{x}$, and inference is done unconditionally with respect to the training set $u = \boldsymbol{x}$. If the assumption of multinormality really holds, efficiency calculations by Efron (1975) lead to the conclusion that one should not condition upon the training set, i.e., choose linear discriminant analysis. For the general case where these assumptions do not hold, none of the procedures seem to dominate.

7 Conditioning and the conditionality principle

Any statistical investigation has to start with a conceptual analysis. This includes choosing question of interest, collect earlier information on this question, the choice of design or sampling plan, choosing target population and sampling units, the choice of a model and maybe of a loss function etc.. The result of this analysis must be considered as a part of the context of the estimation and prediction problem. Then data are collected.

Assume now the setting 1), 2) of the previous section and that u is a (z, τ)-ancillary quasi-observator. All the examples in subsection 6.2 satisfy these conditions.

To choose the conditioning in example 11, one must specify further. Suppose that the data collection is done by first finding out whether the person in question is a man or woman, thereafter asking about cellphone ownership, then the conditioning should be done upon sex. In the opposite case, if the data collection is done from a register of cellphone owners, later asking about sex, then one should condition upon cellphone ownership. In the case where the data are found from a register containing both information on sex and cellphone ownership, one should perhaps condition upon both variables, even though we don't have joint ancillarity here.

The last three examples describe different situations. First consider Example 12. Here the parameter of interest θ is a function of the total parameter $\phi = (\mu_1, \mu_2)$, and an ancillary for ϕ is the choice u of a person to investigate. This is chosen randomly, and is unknown to the experimentalist. Assuming that u takes some definite value, let μ_u be the value of μ for the specific person chosen. The experiment, in whatever way it is done, can then in principle be parametrized by $\phi' = (\mu_u, \theta)$, since the other μ is $2\theta - \mu_u$. From this point of view μ_u is irrelevant for the statistical decision that we want to do. In such a situation one should *not* condition. However, a more difficult situation occurs if one has some independent information about the chosen person which is relevant for θ itself or for the potential estimate of θ. Then it is impossible to obtain a sampling situation. One could or could not condition upon u, but in neither case we do not get any immediate information about θ.

The generalized principle for conditioning (GPC). *Assume that u is a maximal (z, τ)-ancillary quasi-statistic for an e-variable θ.*

1) In the case where u is a statistic, i.e., a known function of z, any inference on θ based upon the data z should be conditional upon u. If there are several maximal such u's to choose between, one should condition upon the one corresponding to the data that have first been obtained.

2) If knowledge of u implies knowledge about a conceptual variable for the observed unit which one is sure is irrelevant for the statistical decision, then examples seem to indicate that one should not condition upon u.

3) The difficult case is when the knowledge of u implies knowledge on some conceptual variable and part of this conceptual variable is relevant for the decision or one is not sure whether or not this is the case. Then one should either seek more information on this conceptual variable or one should perhaps do some suitable model reduction, see below.

Part 1) is consistent with the conditioning chosen in Example 9 and in Example 11. Part 2) is consistent with the decision not to condition in Example 12. It is also applicable to Example 14, and consistent with the results of Efron (1975) for this situation. If one is sure that the underlying distribution for each class is multinormal (with the same covariance matrix), then the parameter of interest is $\theta = \{(\beta_{k0}, \boldsymbol{\beta}_k; k = 1, ..., K-1\}$, but in the LDA case there are underlying additional parameters which are not relevant for the classification. In this case one should not condition, i.e., use LDA instead of logistic regression.

In case we are not sure that the underlying distribution is multinormal, one is in the difficult case 3), and may want to seek more information.

Example 13 is a bordering situation. In most cases, a user of the results of the sensory analysis will not be interested in which assessors that are chosen, will not ask for this information and will thus not condition upon this information.

I have given a normative form of the conditionality principle. For the further development in this book it is also important to consider a descriptive form, which is often given in the literature; see Berger and Wolpert (1988). In this case, the notion 'one shold condition upon ...' translates into '... the unconditioned experiment contains no relevant experimental evidence on θ in addition to that of the conditioned experiment'. As in Berger and Wolpert (1988), the concept of 'relevant experimental evidence' is left undefined, i. e., it can be made precise in any reasonable way.

THE GENERALIZED WEAK CONDITIONALITY PRINCIPLE (GWCP). *Suppose that there are two experiments E_1 and E_2 with common e-variable θ and with equivalent contexts τ. Consider the mixed experiment E^*, whereby $u = 1$ or 2 is observed, each having probability $1/2$ (independent of θ, the data of the experiments and the contexts), and the experiment E_u is then performed. Then the evidence about θ from E^* is just the same as the evidence from the experiment actually performed.*

Note that this corresponds to the situation 1) of the GPC: The variable u is a statistic here; the two experiments are known to the experimentalist. For a note on the equivalence of contexts, see the next section.

The present book concentrates upon estimation and prediction, but the conceptual framework discussed here is also valid for other types of statistical inference. Confidence intervals may be considered if the context contains a set of hypothetical situations where a particular estimation procedure is used, and Bayesian analysis is relevant if prior distributions of parameters are part of the context. Finally, in the case of a hypothesis testing setting, which is not discussed further in this book, the context may contain a specification of a null hypothesis and an alternative hypothesis. Or if we are interested in Fisherian p-value testing, a null hypothesis and a direction for the alternative should be specified in the context.

8 The sufficiency and likelihood principles

The motivation behind the definition of a sufficient statistic is that one wants to reduce the data and still get the same information about the e-variable. One version of the sufficiency principle, as formulated in Berger and Wolpert (1988), translates to our setting as follows:

THE GENERALIZED WEAK SUFFICIENCY PRINCIPLE (GWSP). *Consider an experiment in a context τ as described above, let z be the data of that experiment,*

and let θ be an e-variable connected to the experiment. Assume that 1) and 2) of Section 6 are satisfied. Let $t = t(z)$ be a (z, τ)-sufficient statistic for θ. Then, if $t(z_1) = t(z_2)$, the data z_1 and z_2 contain the same experimental evidence about θ in the context τ.

There can be given many examples to support the GWSP. The simplest example is an independent measurement series $z = (x_1, ..., x_n)$, where the x_i's are normal (μ, σ^2). If σ^2 is known, $\bar{x} = n^{-1} \sum x_i$ is sufficient for μ, and any reasonable inference is based upon $\bar{x}$. If σ^2 is unknown, then $t(z) = (\bar{x}, s^2)$ is sufficient for $\theta = (\mu, \sigma^2)$, where $s^2 = (n-1)^{-1} \sum (x_i - \bar{x})^2$. (The denominator $n-1$ makes s^2 an unbiased estimator of σ^2; more information about this denominator will be given below.) Any reasonable inference on θ under the normal model is based upon $t(z)$. This kind of data reduction was Fisher's motivation for introducing the concept of sufficiency.

Now following an argument from Berger and Wolpert (1988), using the GWSP and the GWCP, which we will regard as more or less obvious, we can derive the following likelihood principle. This result is a classical theorem first given by Birnbaum (1962). The argument is reproduced for completeness in Appendix 2 for the discrete case; this is in fact the case I need later in the discussion of quantum mechanics. For the continuous case, see Berger and Wolpert (1988).

A version of the likelihood principle will be used later to motivate Born's formula in quantum mechanics.

THE GENERALIZED LIKELIHOOD PRINCIPLE. *Consider two experiments with equivalent contexts τ, and assume that θ is the same full e-variable in both experiments. Suppose that two observations z_1^* and z_2^* have proportional likelihoods in the two experiments, where the proportionality constant c is independent of θ.*

Assume that one is sure that the decision problem on θ does not depend any irrelevant UCV. Then these two observations produce the same evidence on θ in this context.

Two contexts τ and τ' are defined to be equivalent if there is a one-to-one correspondence between them: $\tau' = f(\tau)$, $\tau = f^{-1}(\tau')$.

Since both my definition of ancillary and my definition of sufficient statistic depend on the context, and therefore the context is kept fixed in the corresponding principles, it is important that it is kept essentially fixed also here. This aspect makes the generalized likelihood principle weaker than the principle as formulated in the literature, in particular in Berger and Wolpert (1988). On the other hand, paradoxes like what the ordinary likelihood principle seems to imply in the following situation are avoided.

EXAMPLE 15. Suppose that $s_1, s_2, ...$ are independent, identically distributed variables with $P(s = 1) = \theta$ and $P(s = 0) = 1 - \theta$, i.e., iid Bernoulli variables with parameter θ. In experiment $\mathcal{E}_1$, a fixed sample size of 10 observations is

decided upon, and the sufficient statistic $t_1 = \sum_{i=1}^{10} s_i$ turns out to be $t_1 = 8$. In experiment $\mathcal{E}_2$, it is decided to take observations until a total of 2 zeroes has been observed. Then assume that the sufficient statistics $t_2 = \sum s_i$ also turns ot to take the value 8. The two likelihoods are proportional, but the contexts are different, so the intuition that the two experiments may lead to different inference on θ is supported by my version of the likelihood principle. For further discussion of this example, see Berger and Wolpert (1988) and references there.

The introduction of a context makes my formulation of the likelihood principle far less controversial than the ordinary formulation. According to the ordinary principle, the way data are obtained is irrelevant to inference; all information is contained in the likelihood. Thus sampling plans, randomization procedures, and stopping rules are irrelevant according to a common interpretation of the ordinary principle. Furthermore, common frequentist concepts like bias, confidence coefficients, levels and powers of statistical tests, etc., are irrelevant, as they depend on the sample space, not only on the observed observations. In my formulation, all these concepts are related to the context. Also Bayesian priors, if needed, are contained in the context. Maximum likelihood estimation can not be derived from the likelihood principle, but is obviously permissible as a method of obtaining reasonable proposals for estimates in general.

An important special case of the generalized likelihood principle is when the proportionality constant c is equal to 1. Then the two observations z_1^* and z_2^* have equal likelihoods. Again an important special case is when the two experiments are identical. A consequence of the generalized likelihood principle is then that all experimental evidence, given the context, is a function of the likelihood of the experiment.

In the situation 2) of GPC the likelihood principle can not be deduced in a similar way. It is nevertheless clear in the examples how inference should be carried out; in Example 12 by using the sampling distribution, in Example 14 by using the underlying Gaussian distribution. Other examples could be discussed in a similar way. I will assume that the likelihood is the basis for inference also in this case. However, in the situation 3) of GPC there are difficulties in finding the best inference procedure.

9 Estimation and prediction

9.1 Estimation and model reduction

Sufficiency and ancillarity in the case with nuicance parameters have been discussed from many points of view by several authors (Fraser, 1956; Dawid, 1975; Basu, 1977; Godambe, 1980; Zhu and Reid, 1994; Reid, 1995). Here I will see it in light of the assumed context. If θ is the parameter we are interested in, and λ is the rest of the parameters, it may or may not be that λ is relevant for the estimation of θ. In any case, λ and the eventual estimation of λ must be taken

as a part of the context when estimating the parameter of interest θ. Before estimation, λ is an UCV in the context of the inference problem of interest. If this UCV is not relevant, we are in the situation 2) of the GPC, if it is relevant, we are in the more difficult situation 3), and more information should be sought, for instance by estimating λ. However, even when this is possible, it might be better to eliminate λ by reducing the model.

EXAMPLE 8 (CONTINUED). As was seen from the likelihood, σ is irrelevant for the estimation of μ, while μ is of relevance for the estimation of σ. Thus $\widehat{\mu} = \bar{y}$ from any point of view, while the estimation of σ as an isolated parameter of interest may be discussed.

I will promote the REML (restricted or reduced maximal likelihood) principle as a solution to this and similar variance estimation problems. This reduces the likelihood and eliminates the expectation based nuisance parameter if the covariance parameters are the ones of interest:

Let the n-vector $\boldsymbol{y}$ be modelled as multinormal $N(\boldsymbol{X\beta}, \boldsymbol{\Sigma})$, where $\boldsymbol{X}$ is a known $n \times p$ matrix of rank p, and where $\boldsymbol{\Sigma}$ depends on an r-dimensional parameter of interest $\boldsymbol{\theta}$. In general, $\boldsymbol{y}$ is multinormal $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if $\boldsymbol{c}^T\boldsymbol{y}$ is $N(\boldsymbol{c}^T\boldsymbol{\mu}, \boldsymbol{c}^T\boldsymbol{\Sigma c})$ for any constant vector $\boldsymbol{c}$.

Define the residuals $\boldsymbol{r} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{y}$. Let $\boldsymbol{A}$ be an $n \times (n-p)$ matrix of full rank $n-p$ such that $\boldsymbol{A}^T\boldsymbol{X} = \boldsymbol{0}$. Then $\boldsymbol{a} = \boldsymbol{A}^T\boldsymbol{r} = \boldsymbol{A}^T\boldsymbol{y}$ will have a non-singular distribution, and the maximum likelihood estimator found from the distribution of $\boldsymbol{a}$ is independent of the choice of the matrix $\boldsymbol{A}$ with the stated properties. This is called the REML estimator.

EXAMPLE 8 (CONTINUED). REML gives $\widehat{\sigma}^2 = \sum(y_i - \bar{y})^2/(n-1)$ with the correct denominator.

Similarly, REML gives variance components estimates with the correct degrees of freedom (denominator) in all balanced mixed models. In particular, a general problem raised by Neyman and Scott (1948) is solved in a satisfactory way by this estimator. REML was proposed for unbalanced mixed models by Patterson and Thompson (1971), has been discussed by many authors, and is now the routine method when estimating variance components in animal breeding.

In Section 3, I discussed model reduction to one orbit or to a few orbits of a group defined on the e-variable space (parameter space). This may be a way to get rid of nuisance parameters. But there is also a complementary possibility. The orbits of the group G_0 on the sample space constitute equivalence classes there, so we can always index the classes by some a. This is called the maximal invariant of the group. Under weak conditions (see Eaton, 1989, and references there), we can choose a so that it can be given a probability distribution. Similarly, we can index the orbits in the parameter space by some parameter τ, the maximal invariant of the group G. One can easily prove (Lehmann, 1959) that a has a distribution which depends only upon τ, and this gives again a

reduced model. Again, the hope is that in this reduced model, the parameter of interest θ is still present, while the effect of the nuisance parameter λ is reduced or disappears.

As an example of the latter procedure, let once again the n-vector $\boldsymbol{y}$ be modelled as $N(\boldsymbol{X\beta}, \boldsymbol{\Sigma})$, where $\boldsymbol{X}$ is a known $n \times p$ matrix of rank p, and where $\boldsymbol{\Sigma}$ depends on an r-dimensional parameter of interest $\boldsymbol{\theta}$. This model implies that the expectation of $\boldsymbol{y}$ belongs to the vector space V spanned by the columns of $\boldsymbol{X}$. Let the group G_0 consist of all translations of $\boldsymbol{y}$ by vectors in this space. Then the corresponding parameter group G is given by all translations $\boldsymbol{\beta} \to \boldsymbol{\beta} + \boldsymbol{c}$. The orbits of G can be indexed by $\boldsymbol{a}$, where $\boldsymbol{a} = \boldsymbol{A}^T\boldsymbol{r} = \boldsymbol{A}^T\boldsymbol{y}$, with $\boldsymbol{A}$ being any $n \times (n-p)$ matrix of full rank $n-p$ such that $\boldsymbol{A}^T\boldsymbol{X} = \boldsymbol{0}$, and where $\boldsymbol{r}$ are the residuals of the model. The orbits of G in the parameter space are independent of the nuisance parameter $\boldsymbol{\beta}$ and depend only upon the parameter of interest $\boldsymbol{\theta}$. Thus this model reduction gives exactly what one wants, and maximum likelihood estimation in the reduced model in just REML.

9.2 Prediction, sufficiency and partial least squares

Consider now the prediction setting described at the end of Section 5. We will let y_i be scalars with identical independent conditional distributions, given vectors or scalars x_i and perhaps a distribution if the x_i's. The conditional distribution of y_i, given x_i depend on a vector or scalar parameter θ for $i = 0, 1, 2, ..., n$. The training set consists of the observations for $i = 1, ..., n$, which are OCV's. In addition, we know x_0, and want to predict y_0.

For later discussions of the link to quantum theory, note that the prediction problem per se is dependent upon the e-variable y_0 connected to a single item, not to a population. The likelihood, given this e-variable is the conditional density of x_0, given y_0. This conditional distribution played a prominent rôle in the discussion of Cook (2007).

Having noted this, this prediction problem is obviously dependent upon the UCV θ. The first step now of the prediction procedure is to determine which part of θ is relevant for the prediction. In the linear regression case this is just the regression vector, or more specifically $\boldsymbol{\beta}^T\boldsymbol{x}_0$, where $\boldsymbol{\beta}$ is the regression vector and $\boldsymbol{x}_0$ is now seen as a vector. The second step is to estimate this part, eventually after a model reduction. In Hastie et al. (2009) a large number of estimation procedures are described for the linear regression case, and also for other cases. In Example 6 of Section 3, we concentrated on a method to reduce the model before doing the estimation.

Model reduction, in regression models as well as in other models can be motivated from many points of view. Cook (2007) related this to data reduction and considered regression of a random variable y in with respect to a random vector $\boldsymbol{x}$. A reduction R in $\boldsymbol{x}$ from p dimensions to m dimensions, $m < p$ was

said to hold if one of the following equivalent statements holds:

(i) $\boldsymbol{x}|(y, R(\boldsymbol{x}))$ has the same distribution as $\boldsymbol{x}|R(\boldsymbol{x})$,
(ii) $y|\boldsymbol{x}$ has the same distribution as $y|R(\boldsymbol{x})$,
(iii) y is independent of $\boldsymbol{x}$ given $R(\boldsymbol{x})$.

Cook (2007) states that this is quite analoguous to the ordinary definition of sufficiency. In fact it is equivalent to Definition 4 above for the UCV $\theta = y$ when the data are given by $\boldsymbol{z} = \boldsymbol{x}$.

A special case is when the reduction is linear: $R(\boldsymbol{x}) = \boldsymbol{P}\boldsymbol{x}$ for some projection operator $\boldsymbol{P}$ of rank m. Then (iii) is equivalent to

(iv) y is independent of $\boldsymbol{Q}\boldsymbol{x}$ given $\boldsymbol{P}\boldsymbol{x}$,

where $\boldsymbol{Q} = \boldsymbol{P}^{\perp}$.

Consider now the case where $\boldsymbol{x}$ is multinormal. Then a stronger statement is obtained if we add the extra condition

(v) $\boldsymbol{Q}\boldsymbol{x}$ is independent of $\boldsymbol{P}\boldsymbol{x}$,

and this is equivalent to $\mathrm{Cov}(\boldsymbol{Q}\boldsymbol{x}, \boldsymbol{P}\boldsymbol{x}) = \boldsymbol{0}$. It is proved in Cook et al. (2012) that in the multinormal case (or in general when independence is weakened to uncorrelatedness) the statement (iv)+(v) is equivalent to the statement that the reduced model is an envelope model of dimension m (see Cook et al. (2010) for a definition and a comprehensive discussion). Furthermore, it is proved in op. cit. that this envelope model is equivalent to the population PLS model with m relevant components, that is, the same model that was motivated in Section 3 from invariance.

Parameters in the reduced model are estimated in op. cit. by maximum likelihood and other methods. Bayesian estimation is discussed in Helland et al. (2012).

In the next part of the book I will address quantum mechanics from an epistemic point of view. A crucial concept is then that of an inaccessible e-variable, that is, an epistemic variable which cannot be estimated with arbitrary accuracy by any experiment. In the regression model where the dimension p by necessity is larger than the number n of observations, the regression vector $\boldsymbol{\beta}$ must be seen as an inaccessible parameter. However, under suitable circumstances, the e-variable function $\boldsymbol{\beta}^T\boldsymbol{x}_0$ may still be estimable. In particular this is the case when $\boldsymbol{x}_0$ is regarded as random, with the same distribution as the other $\boldsymbol{x}_i$'s. One approach towards estimating this function may be model reduction as discussed in this section.

PART II

10 Inaccessible conceptual variables and quantum theory

The statistical literature is full of discussions on how to do inference, but contains very little on the choice of question to do inference on in some given situation. These different questions may be conflicting, even complementary. In the following sections I will start by formalizing a way in which the discussion of such complementary questions may be addressed in the extreme case where it is only possible to raise one out of many different possible questions at a time. Each such question will be an epistemic question 'What is θ?' for some e-variable θ, and I will assume that the epistemic process ends by giving some information about θ, in the simplest case a complete specification: $\theta = u_k$.

The concept of an epistemic process is taken to be very wide in this book. In addition to statistical questions concerning a parameter θ, we can think of questions like: How many sun hours will there be here tomorrow? At the outset, to address this epistemic question will involve meteorological expertise and a lot of data from similar situations, but tomorrow the question can be answered by just counting the number of sun hours. Both these processes will be seen as epistemic processes.

So far I have assumed that each relevant e-variable is accessible, that is, it can be assessed with arbitrary accuracy by some experiment. In Helland (2006, 2008, 2010) several situations with inaccessible conceptual variables were described, and it was indicated that such situations in special cases could form a link to important parts of quantum theory. I consider this way of thinking to be essential as a step towards obtaining a unification of epistemic science, and also as an attempt to give an alternative background for the - from a statistical point of view and also from the layman's point of view - very formal language that one finds in textbooks and in scientific publications, both within quantum physics and in the mathematical traditions developed from this. In the following sections a less formal approach will be presented. Compared to my earlier publications, the discussion here will hopefully give both a simpler and a more complete treatment of my approach towards quantum mechanics.

In statistics, the parameter concept is connected to a hypothetical population of items. My e-variables are intended also for situations where we have a single item or a few items, and a human subject or a group of subjects use these variables in making statements about the item(s). This is crucial for my epistemic interpretation of quantum mechanics, an interpretation which I also share with the Bayesian quantum foundation school; see below.

Quantum theory has a long history starting with the work of several eminent physicists in the beginning of the previous century, via the formalization made by von Neumann (1932) to the rather intense debate on quantum foundation that we see today. Interpretations of the theory have been given by several

authors, but it has also been argued that no interpretation is needed; see Fuchs and Peres (2000). Several attempts have been made recently to derive quantum theory from a few explicit or implicit physical assumptions; see Hardy (2001), Chiribella et al. (2010), Masanes (2010), Fields (2011) and Fivel (2012). There is also a group of quantum foundation researchers working towards a link with Bayesian inference; see Caves et al. (2002), Schack (2006), Timpson (2008), Fuchs (2010) and Fuchs and Schack (2011). The use of quantum information theory in the exploration of the foundation has also recently proved to be very useful, see Fuchs (2002). The present work has much in common with these schools, but I find it fruitful to maintain a broader link to statistics, in particular to allow a broader view on statistical inference than just the Bayesian view. In this way I will argue for a foundation which is purely epistemological: A general approach for going from experienced data to information about the nature behind these data.

One very obvious case of an inaccessible conceptual variable is in connection to counterfactual reasoning. Assume a single medical patient and let the doctors have the choice between two mutually exclusive treatments. Let θ^i be the time for this patient until recovery when treatment i is used ($i = 1, 2$), and let $\phi = (\theta^1, \theta^2)$. Then θ^1 or θ^2 can be predicted before the treatment is applied, and each of them can be determined precisely after some time period, but ϕ is inaccessible, that is, there is no procedure by which ϕ can be assessed with arbitrary accuracy at any time for a single patient by any medical doctor, by any scientist or by any observer. This can be amended by instead of one patient considering large homogeneous groups of patients, which is done in standard statistical texts, but in practice there is a limitation on how homogeneous a group of patients can be. And concepts may be of interest for one single patient, too.

Here are two other examples of inaccessible conceptual variables:

- We want to measure some quantity θ^1 with a very accurate apparatus which is so fragile that it is destroyed after a single measurement. There is another quantity θ^2 which can only be found by dismantling the apparatus, and then it can not be repaired. The vector $\phi = (\theta^1, \theta^2)$ is again inaccessible.

- Assume that two questions are to be asked to a single individual at some given moment, and that we know that the answer will depend on the order in which the questions are posed. Let the e-variable (θ^1, θ^2) be the answers when the questions are posed in one order, and let the answers be (θ^3, θ^4) when the questions are posed in the opposite order. Then the vector $\phi = (\theta^1, \theta^2, \theta^3, \theta^4)$ is inaccessible.

From a statistical point of view: Inaccessible parameters also occur in linear models of non-full rank, often used in the case of unbalanced data, cp. Searle (1971), and in the analysis of designed experiments where only some contrasts can be estimated. Also, in regression models where the number of variables by necessity is larger than the number of observations, the regression parameter is

an inaccessible parameter. In my opinion a more complete theory of statistical inference is definitely obtained if we allow for inaccessible conceptual variables.

It is a crucial fact that the inaccessible conceptual variables take abstract values in some mathematical space and that operations such as group actions may be made on this space. This is the case with the counterfactual example above, where a group action such as a change of time scale can be made. However, I will not regard the inaccessible conceptual variables as physical variables, and they do not take concrete values, so I am not developing a hidden variable theory of the kind that has been much debated in the physical literature over the years.

An example of a hidden variable theory is David Bohm's dual wave-particle theory, and John Bell (see Bell, 1987) proved that this theory is non-local. In fact, Bell proved much more. His famous theorem states that any realistic theory consistent with quantum mechanics must be non-local. This result has been very important in discussions among physicists in recent years. Bell's theorem is proved using the socalled Einstein-Podolski-Rosen experiment and Bell's inequality, concepts which for completeness will be discussed later in this book. One point for me here is that I do not want to develop a non-local theory, that is, a theory where communication is made by signals travelling faster than the light speed. Then I am instead forced to take a closer look upon the concept of realism. This has also been done recently in a very convincing way by Nisticò and Sestito (2011). In that paper they take as a point of departure the criterion of reality as formulated be Einstein et al. (1935):

CRITERION OF REALITY. *If, without in any way disturbing a system, we can predict the value of a physical quantity, then there exists an element of physical reality corresponding to this physical quantity.*

Following arguments from Bohr's discussion of Einstein et al. (1935) they make the case for a strict interpretation of this criterion:

STRICT INTERPRETATION. *To ascribe reality to P, the measurement of an observable whose outcome would allow for the prediction of P, must actually be performed.*

Nisticò and Sestito (2011) go on and formulate an extension of quantum correlation which is consistent with the strict interpretation, and using this they show that Bell's argument and several related arguments in the literature fail when realism is interpreted in this strict way. Thus the possibility turns out to be open to interpret the non-locality theorems in the physical literature as arguments supporting the strict criterion of reality, rather than as a violation of locality.

Since the present book is theoretical and not experimental, I will have to modify Nisticò and Sestito's requirement of strict interpretation slightly: '... a description of how the measurement can be actually performed, must be given.' It is important that my conceptual variables are thought of as defined by one

person or a group of persons and to the experimental data that he/she/they are able to obtain.

In other papers, Bell's theorem is interpreted as saying that quantum physics must necessarily violate either the principle of locality or counterfactual definiteness. Counterfactual definiteness is defined as the ability to speak with meaning of definiteness of results of measurements that have not been performed (i.e., the ability to assure the existence of objects, and properties of objects, even when they have not been measured.) In this book it is crucial that I do not assume counterfactual definiteness. All my conceptual variables are assumed to be defined by some person(s), and these conceptual variables will not necessarily be such that results of measurements not performed will have meaning. Here is a simple example: By first sight, one of the statements 'I have something on my lap' and 'I do not have anything on my lap' must be true. But if I am standing, neither of these statements are true. The logical status of statements must depend on the context.

In my formulation, I will look upon the accessible e-variables as variables connected with experiment which actually can be imagined to be performed by some person. This person will have a certain context for his experiment. It is possible that another person, who has no communication with the first one, has a different context and uses different e-variables to formulate his observations, therefore getting seemingly conflicting predictions. But as soon as communication is restored, there must be no conflict any more. To make this precise: The two persons must then make non-conflicting predictions if they agree on a common context, and they must agree on observed results as long as they both have observed results.

11 The maximal symmetrical epistemic setting; definitions

I proceed to discuss a setting from which I will show that essential parts of the formalism of quantum mechanics can be derived. From my point of view this is nothing but a special situation with an inaccessible conceptual variable, where I focus upon accessible sub conceptual variables and where symmetry is introduced by natural group actions. The purpose at this point is not to derive all aspects of quantum mechanics, only as much that we see that the e-variable concept is useful also in this connection, so that we can obtain an interpretation where there is a link to the ordinary statistical theory of estimation/prediction. Later the assumptions made here will be weakened.

Let in general ϕ be an inaccessible conceptual variable taking values in some topological space Φ, and let $\lambda^a = \lambda^a(\phi)$ be accessible functions for a belonging to some index set $\mathcal{A}$. I will repeat that an e-variable is accessible if it in the given context can be estimated with arbitrary accuracy by *some* experiment. Technically I will without further mention assume that all functions defined on Φ are Borel-measurable. To begin with, I will assume that the functions λ^a are

maximal, and also that there is an isomorphism between them.

ASSUMPTION 1. *a) Consider the partial ordering defined by $\alpha < \beta$ iff $\alpha = f(\beta)$ for some function f. Under this partial ordering each $\lambda^a(\phi)$ is maximally accessible.*

b) For $a \neq b$ there is an invertible transformation g_{ab} such that $\lambda^b(\phi) = \lambda^a(g_{ab}(\phi))$.

Note that the partial ordering in a) is consistent with accessibility: If β is accessible and $\alpha = f(\beta)$, then α is accessible. Also, ϕ is an upper bound under this partial ordering. The existence of maximal accessible conceptual variables follows then from Zorn's lemma.

Below, I will often single out a particular index $0 \in \mathcal{A}$. Then a), given b), can be formally weakened to the assumption that $\lambda^0(\phi)$ is maximally accessible, and b) can be weakened to the existence for all a of an invertible transformation g_{0a} such that $\lambda^a(\phi) = \lambda^0(g_{0a}(\phi))$. Take $g_{ab} = g_{0a}^{-1} g_{0b}$.

In the example above with counterfactual medical treatments, we can take $\lambda^a = \theta^1$, $\lambda^b = \theta^2$, $\phi = (\lambda^a, \lambda^b)$ and $g_{ab}((\lambda^a, \lambda^b)) = (\lambda^b, \lambda^a)$. In general, when the transformation of Assumption 1b exists, it is usually easy to see how it can be chosen.

Even though ϕ is inaccessible, it is possible to operate on ϕ with functions, in particular group actions. For instance, in the medical example above, one can operate on ϕ with a common change of time unit. It is then important to make sure that these actions induce unique operations on the accessible e-variables λ^a. The property which ensures this, is given by:

DEFINITION 8. *Let a group H act upon a conceptual variable ϕ, and let $\eta = \eta(\phi)$ be a sub conceptual variable. Then η is said to be permissible with respect to H if $\eta(\phi_1) = \eta(\phi_2)$ implies $\eta(h\phi_1) = \eta(h\phi_2)$ for all $h \in H$.*

When η is permissible with respect to H, one can define a group $\tilde{H}$ of actions upon η by $\tilde{h}\eta(\phi) = \eta(h\phi)$. For a group $\tilde{H}$ acting upon η one can always find at least one corresponding group H acting upon ϕ.

For a given η, there is a unique maximal group with respect to which η is permissible. This is the group of actions h for which $\eta(\phi_1) = \eta(\phi_2)$ is equivalent to $\eta(h\phi_1) = \eta(h\phi_2)$ for all pairs (ϕ_1, ϕ_2).

Let us now go back to the situation of Assumption 1. We single out one particular index $0 \in \mathcal{A}$.

DEFINITION 9. *a) Let G^0 be the maximal group of transformations of Φ under which $\lambda^0(\phi)$ is permissible, and let $G^a = g_{0a}^{-1} G^0 g_{0a}$.*

b) Let G be the group generated by G^0 and the transformations g_{0a}.

It is easy to see that G^a is the maximal group under which λ^a is permissible and that G is the group generated by $G^a; a \in \mathcal{A}$ and the transformations g_{ab}. In this setting I want to introduce the further

ASSUMPTION 2. *a) The group G is a locally compact topological group, and satisfies week assumptions such that an invariant measure on Φ exists. (see Appendix 3).*

b) $\lambda^a(\phi)$ varies over an orbit or a set of orbits of the smaller group G^a. This is made precise in the following way: λ^a varies over an orbit or a set of orbits of the corresponding group $\tilde{G}^a$.

c) The group generated by products of elements of $G^a, G^b, ...; a, b, ... \in \mathcal{A}$ is equal to G.

Assumption 2a) is a technical one, needed in the next section. Note that G is defined in terms of transformations upon Φ, so that the topology must be introduced in terms of these transformations. Technically this can be achieved by assuming Φ to be a metric space with metric d, and letting $g_n \to g$ if $\sup_\phi d(g_n(\phi), g(\phi)) \to 0$. Assumption 2b) can be motivated from Example 16 below. Concerning 2c), it follows from $g^a g^b ... = g_{0a}^{-1} g^0 g_{0a} g_{0b}^{-1} g^{0'} g_{0b}, ...$, where $g^a \in G^a, g^b \in G^b, ...$ and $g^0, g^{0'}, ... \in G^0$, that the group of products is contained in G. That it is equal to G, is an assumption on the richness of the index set $\mathcal{A}$ or the richness of G^0.

The setting described here, where Assumption 1 and Assumption 2 are satisfied, includes many quantum mechanical situations including spins and systems of spins. I will call it *the maximal symmetrical epistemic setting*. Later I will also sketch a macroscopical situation where the assumptions of the maximal symmetrical epistemic setting are satisfied. I hope to discuss this latter subject further elsewhere, but the focus in the present book will be quantum-mechanical.

EXAMPLE 16. Model the spin vector of a particle such as the electron by a vector ϕ, an inaccessible conceptual variable. More generally, we can let ϕ denote the total spin/angular momentum vector for any particle or system of particles. Let the symmetry group G be the group of rotations of the vector ϕ, that is, the group that fixes the norm $\|\phi\|$. Next, choose a direction a in space, and focus upon the spin component in this direction:

$$\zeta^a = \|\phi\| \cos(\phi, a).$$

The largest subgroup G^a with respect to which ζ^a is permissible, is given by rotations around a together with a reflection in a plane perpendicular to a. However, the action of the group $\tilde{G}^a$ on ζ^a is just a reflection together with the identity.

Finally, introduce model reduction of the kind discussed in Section 3: The orbits of $\tilde{G}^a$ as acting on ζ^a are given as two-point sets $\{\pm c\}$ together with the single point 0. A maximal model reduction is to one such orbit. Later I will give arguments to the effect that we want to reduce to the a set of orbits indexed by an integer or half-integer j, and that we will let this reduced set of orbits be

$$-j, -j+1, ..., j-1, j.$$

Letting λ^a be the parameter ζ^a reduced to this set of orbits of G^a, and assuming this to be the maximal accessible parameter, we can prove that the general assumptions of this section are satisfied (except in the case $j = 0$, where we must redefine $G = G^0$ to be the trivial group). For instance, here is an indication of an argument leading to the proof of Assumption 2c): Given a and b, a transformation g_{ab} sending $\lambda_a(\phi)$ onto $\lambda_b(\phi)$ can be obtained by a reflection in a plane P perpendicular to a plane containing the two vectors a and b, where P contains the mid-line between a and b.

The case with one orbit and $c = 1/2$ correspond to electrons and other spin $1/2$ particles. The direction defined by $a = 0$ is some arbitrary fixed direction.

In general, Assumption 2b) may be motivated in a similar manner: First, an e-variable ζ^a is introduced for each a through a chosen focusing; then define G^a as the maximal group under which ζ^a is permissible, and finally λ^a as a reduction of ζ^a to a set of orbits of $\tilde{G}^a$. The content of Assumption 2b) together with Assumption 1 is that it is *this* λ^a which is maximally accessible. This may be regarded as the quantum hypothesis.

12 The maximal symmetrical epistemic setting; Hilbert space

The crucial step now towards the formalism of quantum mechanics is to define a Hilbert space, that is, a complete inner product space which serves as a state space in the formalism (see Appendix 3). In ordinary quantum mechanics all observables are identified with operators on such a Hilbert space and every state is identified with a unit vector in the Hilbert space or more generally with a ray proportional to a unit vector. There is a large, fairly abstract general theory on this, well known to physicists, but largely unknown to statisticians. My goal here is to rederive this theory from the assumptions of the maximal symmetrical epistemic setting. This may serve as introducing statisticians and other professionals to the theory, and also serve as a link between epistemic cultures. This section is somewhat technical, and can be skimmed at the first reading, but it is essential for what I feel should be the way to understand ordinary quantum theory.

12.1 A preliminary solution

By Assumption 2a) there exists an invariant measure ρ for the group's action:

$$\rho(gA) = \rho(A)$$

for all $g \in G$ and for all Borel-measurable subsets A of Φ. In general there is a distinction between right-invariant and left-invariant measures (again see Appendix 3), but I will limit myself here to compact groups and other situations where the two measures coincide. This is not crucial, however. There are

general arguments in Helland (2010) that ρ always should be chosen as the right-invariant measure. If G is transitive on Φ, then ρ is unique up to a multiplicative constant. For compact groups, ρ can be normalized, i.e., taken as a probability measure. In the case of a compact group with r orbits, we take $\rho = r^{-1}\sum \rho_i$, where ρ_i is the unique invariant normed measure on orbit i.

The measure ρ allows us to define $L^2(\Phi, \rho)$ as the space of all complex measurable functions f for which $\int_\Phi |f(\phi)|^2 \rho(d\phi) < \infty$, equiped with the scalar product $(f_1, f_2) = \int_\Phi f_1^*(\phi) f_2(\phi) \rho(d\phi)$, where f^* denotes complex conjugate, in particular, $\|f\|^2 = (f, f)$. We identify f_1 and f_2 when $\|f_1 - f_2\| = 0$. This then gives a Hilbert space. The following closed subspace is also a Hilbert space:

DEFINITION 10. *In the symmetrical epistemic setting the basic Hilbert space is given by*

$$H = L(\Lambda^0) = \{f \in L^2(\Phi, \rho) : f(\phi) = r(\lambda^0(\phi)) \text{ for some } r\}.$$

Thus H is defined as the set of L^2-functions that are functions of $\lambda^0(\phi)$. In an attempt to link this to the other λ^a's, we first define the (left) regular representation U of the group G: For given $f \in L^2(\Phi, \rho)$ and given $g \in G$ we define a new function $U(g)f$ by

$$U(g)f(\phi) = f(g^{-1}\phi). \tag{6}$$

Without proof I mention 5 properties of the set of operators $U(g)$:
- $U(g)$ is linear: $U(g)(a_1 f_1 + a_2 f_2) = a_1 U(g) f_1 + a_2 U(g) f_2$.
- $U(g)$ is unitary: $(U(g)f_1, f_2) = (f_1, U(g)^{-1} f_2)$.
- $U(g)$ is bounded: $\sup_{f:\|f\|=1} \|U(g)f\| = 1 < \infty$.
- $U(\cdot)$ is continuous: If $g_n \to g_0$ in the group topology, then $U(g_n) \to U(g_0)$ (in the matrix norm in the finite-dimensional case, which is what we will focus on below; in general in the topology of bounded linear operators).
- $U(\cdot)$ is a homomorphism: $U(g_1 g_2) = U(g_1)U(g_2)$ and $U(e) = I$ for the unit element.

The concept of *homomorphism* will be crucial in this section. In general, a homomorphism is a mapping $k \to k'$ between groups K and K' such that $k_1 k_2 \to k_1' k_2'$ when $k_1 \to k_1'$ and $k_2 \to k_2'$, and such that $e \to e'$ for the unit elements. Then also $k^{-1} \to (k')^{-1}$ when $k \to k'$.

A *representation* of a group K is a continuous homomorphism from K into a group of linear operators on some vector space. If the vector space is finite dimensional, the linear operators can be taken as matrices (see Appendix 3). There is a large and useful mathematical theory about operator (matrix) representation of groups; some of it is sketched in Appendix 3. Equation (6) gives one such representation of the basic group G on the vector space $L^2(\Phi, \rho)$.

PROPOSITION 2. *Let* $L(\Lambda^a) = \{f \in L^2(\Phi, \rho) : f(\phi) = r(\lambda^a(\phi)) \text{ for some } r\}$ *and let* $U_a = U(g_{0a})$. *Then*

$$L(\Lambda^a) = U_a^{-1} H \text{ through } r(\lambda^a(\phi)) = U_a^{-1} r(\lambda^0(\phi)).$$

Proof.
If $f \in L(\Lambda^a)$, then $f(\phi) = r(\lambda^a(\phi)) = r(\lambda^0(g_{0a}\phi)) = U(g_{0a})^{-1}r(\lambda^0(\phi)) = U_a^{-1}f_0(\phi)$, where $f_0 \in H$.

Since $a = 0$ is an arbitrary, but fixed index, this gives in principle a unitary connection between the different choices of H, different representations of the 'Hilbert space apparatus'. As already stated, in conventional quantum theory, observables are represented by operators on such a Hilbert space. One of our points is that this formal theory quite generally can be understood in terms of conceptual variables. In principle one could imagine that one could represent everything in the single Hilbert space H through the unitary transformation of Proposition 2.

This simple solution is not satisfactory, however. To see this, look at the discrete case. Then by a reasonable epistemic definition, a state is given by the statement of the form $\lambda^a = u_k^a$, that is, a maximally accessible e-variable λ^a has been chosen, an epistemic question: 'What is λ^a?' has been asked, and after an epistemic process a definite answer is found. In $L(\Lambda^a)$ this can be represented by the indicator function $f_a^k(\phi) = I(\lambda^a(\phi) = u_k^a)$. When transformed into a function in H by Proposition 2, this turns out to be

$$U_a f_k^a = U(g_{0a})I(\lambda^0(g_{0a}(\phi)) = u_k) = U(g_{0a})U(g_{0a})^{-1}I(\lambda^0(\phi) = u_k) = f_k^0.$$

Thus by this simple transformation the indicator functions in H are not able to distinguish between the different questions asked.

12.2 Towards the final solution

Another reason why the simple solution is not satisfactory is that the regular representation U will not typically be a representation of the whole group G on the Hilbert space H. This can however be amended by the following theorem. Its proof and the resulting discussion below are where the Assumption 2c) of the maximal symmetrical epistemic setting is used. Recall that throughout, upper indices (G^a, g^a) are for the subgroups of G connected to the accessible variables λ^a, similarly $(\tilde{G}^a, \tilde{g}^a)$ for the group (elements) acting upon λ^a. Lower indices (e.g., $(U_a = U(g_{0a}))$ are related to the transformations between these variables.

THEOREM 1. *(i) A representation (possibly multivalued) V of the whole group G on H can always be found.*
(ii) For $g^a \in G^a$ we have $V(g^a) = U_a U(g^a) U_a^\dagger$.

The proof of Theorem 1 is given i Appendix 4.

What is meant by a multivalued representation? As an example, consider the group $SU(2)$ of unitary 2×2 matrices. Many books in group theory will

state that there is a homomorphism from $SU(2)$ to the group $SO(3)$ of real 3-dimensional rotations, where the kernel of the homomorphism is $\pm I$. This latter statement means that both $+I$ and $-I$ are mapped into the identity rotation by the homomorphism. For an explicit way to formulate the homomorphism $SU(2) \to SO(3)$, see for instance Knapp (1986), Exercise 5(a) of Chapter 1.

In this case there is no unique inverse $SO(3) \to SU(2)$, but nevertheless we may say informally that there is a multivalued homomorphism from $SO(3)$ to $SU(2)$. Examples of such a discussion can be found in Ma (2007). Here is a way to make this precise:

Extend $SU(2)$ to a new group with elements (g, k), where $g \in SU(2)$ and k is an element of the group $K = \{\pm 1\}$ with the natural multiplication. The multiplication in this extended group is defined by $(g_1, k_1)\cdot(g_2, k_2) = (g_1 g_2, k_1 k_2)$, and the inverse by $(g, k)^{-1} = (g^{-1}, k^{-1})$. Then there is an invertible homomorphism between this extended group and $SO(3)$.

A similar construction can be made with the representation V of Theorem 1.

THEOREM 2. *(i) There is an extended group G' such that V is a univariate representation of G' on H.*

(ii) There is a unique mapping $G' \to G$, denoted by $g' \to g$, such that $V(g') = V(g)$. This mapping is a homomorphism.

THEOREM 3. *(i) For $g' \in G'$ there is a unique $g^0 \in G^0$ such that $V(g') = U(g^0)$. The mapping $g' \to g^0$ is a homomorphism.*

(ii) If $g' \to g^0$ by the homomorphism of (i), and $g' \neq e'$ in G', then $g^0 \neq e$ in G^0.

The proofs of Theorem 2 and Theorem 3 are given in Appendix 4.

Note that while G is a group of transformations on Φ, the extended group G' must be considered as an abstract group.

12.3 The discrete case

In much of this book I will limit myself to the case where the accessible e-variables λ have a finite discrete range. This is often done in elementary quantum theory texts, in fact also in recent quantum foundation papers, and in our situation it has several advantages:

- It is easy to interprete the principle that λ can be estimated with any fixed accuracy.
- In particular, confidence regions and credibility regions for an accessible e-variable can be taken as single points if observations are accurate enough.
- The operators involved (see later) will be much simpler and are defined everywhere.

- The operators can be understood directly from an epistemic setting, see below.

Consider now statements of the form: $\lambda^a = u_k^a$. We start with the following remark: It is possible to simplify the notation for the values taken by λ^a under the maximal symmetrical epistemic setting.

PROPOSITION 3. *The values u_k^a can always be arranged such that $u_k^a = u_k$ is the same for each a ($k = 1, 2, ...$).*

Proof.
By Assumption 1

$$\{\phi : \lambda^b(\phi) = u_k^b\} = \{\phi : \lambda^a(g_{ab}(\phi)) = u_k^b\} = g_{ba}(\{\phi : \lambda^a(\phi) = u_k^b\}).$$

The sets in brackets on the lefthand side here are disjoint with union Φ. But then the sets in brackets on the righthand side are disjoint with union $g_{ab}(\Phi) = \Phi$, and this implies that $\{u_k^b\}$ gives all possible values of λ^a.

So look at the statement $\lambda^a(\phi) = u_k$. This means two things: 1) One has sought information about the value of the maximally accessible e-variable λ^a, that is, asked the question: What is the value of λ^a? 2) One has obtained the answer $\lambda^a = u_k$. This information can be thought of as a perfect measurement, and it can be represented by the indicator function $f_k^a(\phi) = I(\lambda^a(\phi) = u_k)$, which is a function in $L(\Lambda^a)$. From Proposition 2, this function can by a unitary transformation be represented in H, which now is a vector space with a discrete basis, a finite-dimensional vector space: $U_a f_k^a$. However, we have just seen that this tentative state definition $U_a I(\lambda^a(\phi) = u_k) = U(g_{0a}) I(\lambda^0(g_{0a}\phi) = u_k)$ led to ambiguities. These ambiguities can be removed by replacing the two g_{0a}'s here in effect by different elements g'_{0ai} of the extended group G'. Let g'_{0a1} and g'_{0a2} be two different such elements where both $g'_{0a1} \to g_{0a}$ and $g'_{0a2} \to g_{0a}$ according to Theorem 2 (ii). I will prove in a moment that this is in fact always possible when $g_{0a} \neq e$. Let $g'_a = (g'_{0a1})^{-1} g'_{0a2}$, and define what physicists call a ket vector by

$$|a; k\rangle = V(g'_a) U_a I(\lambda^a(\phi) = u_k) = V(g'_a)|0; k\rangle,$$

where $|0; k\rangle = I(\lambda^0(\phi) = u_k)$.

PROPOSITION 4. *Suppose that $\tilde{G}^0$ is transitive on the range of λ^0. Then for each a and k there is a $g'(a, k) \in G'$ such that $|a; k\rangle = V(g'(a, k))|0; 0\rangle$.*

For the proof, see again Appendix 4.

In the following I will not make use of Proposition 4, so I will not need the assumption that $\tilde{G}^0$ is transitive on the range of λ^0. This is fortunate, for in Example 16 this assumption is only satisfied for the spin 0 and the spin 1/2 case ($c = 0$ or $c = 1/2$).

DEFINITION 9. *$|a;k\rangle$ is the state vector in H corresponding to the statement $\lambda^a(\phi) = u_k$.*

INTERPRETATION OF THE STATE VECTOR $|a;k\rangle$: *1) The question: 'What is the value of λ^a?' has been posed. 2) We have obtained the answer $\lambda^a = u_k$. Both the question and the answer are contained in the state vector.*

In order that this interpretation shall make sense, I need the following result, which is proved in Appendix 4. I assume that $\tilde{G}^0$ is non-trivial.

THEOREM 4. *a) Assume that two vectors in H satisfy $|a;i\rangle = |b;j\rangle$, where $|a;i\rangle$ corresponds to $\lambda^a = u_i$ for one perfect measurement and $|b;j\rangle$ corresponds to $\lambda^b = u_j$ for another perfect measurement. Then there is a one-to-one function F such that $\lambda^b = F(\lambda^a)$ and $u_j = F(u_i)$. On the other hand, if $\lambda^b = F(\lambda^a)$ and $u_j = F(u_i)$ for such a function F, then $|a;i\rangle = |b;j\rangle$.*

b) Each $|a;k\rangle$ corresponds to only one $\{\lambda^a, u_k\}$ pair except possibly for a simultaneous one-to-one transformation of this pair.

COROLLARY. *The group G is properly contained in G', so the representation V of Theorem 1 is really multivalued.*

Proof of the corollary.
If we had $G' = G$, then $|a;k\rangle$ and $|b;k\rangle$ both reduce to $U_a I(\lambda^a(\phi) = u_k) = U_b I(\lambda^b(\phi) = u_k) = I(\lambda^0 = u_k)$, so Theorem 4 and its proof could not be valid.

Theorem 4 and its corollary are also valid in the situation where we are interested in just two accessible variables λ^a and λ^b, which might as well be called λ^0 and λ^a. We can then provisionally let the group G be generated by g_{0a}, $g_{a0} = g_{0a}^{-1}$ and all elements g^0 and g^a. The earlier statement that it is always possible to find two *different* elements g'_{0a1} and g'_{0a2} in G' which are mapped onto g_{0a} follows.

Finally we have

THEOREM 5. *For each $a \in \mathcal{A}$, the vectors $\{|a;k\rangle; k = 1, 2, ...\}$ form an orthonormal basis for H.*

Proof.
Taking the invariant measure ρ on H as normalized to 1, the indicator functions $|0;k\rangle = I(\lambda^0(\phi) = u_k)$ form an orthonormal basis for H. Since the mapping $|0;k\rangle \to |a;k\rangle$ is unitary, the Theorem follows.

So if $b \neq a$ and k is fixed, there are complex constants c_{ki} such that $|b;k\rangle = \sum_i c_{ki}|a;i\rangle$. This opens for the interference effects that one sees discussed in quantum mechanical texts. In particular $|a;k\rangle = \sum_i d_{ki}|0;i\rangle$ for some constants d_{ki}. This is the first instance of something that we also will meet later in

different situations: New states in H are found by taking linear combinations of a basic set of state vectors.

In the case of a finite-dimensional space H, the ket vector $|a;k\rangle$ may be looked upon as a coloumn vector. The corresponding dual vector, its complex conjugate row vector is called a bra vector $\langle a;k|$. The scalar product $\langle a;i|\cdot|b;j\rangle$ is written as a bracket $\langle a;i|b;j\rangle$. It follows from Theorem 5 that $\langle a;i|a;j\rangle = \delta_{ij}$, and thus the ket vectors have the norm 1. For any operator A on H we also define the complex scalar $\langle a;i|A|b;j\rangle$.

The information contained in the ket $|a;k\rangle$ is by definition the same as the information contained in the one-dimensional projector $|a;k\rangle\langle a;k|$, where we in general define $|\alpha\rangle\langle\beta|$ by $(|\alpha\rangle\langle\beta|)|\gamma\rangle = |\alpha\rangle\langle\beta|\gamma\rangle$ for all kets $|\gamma\rangle$. In particular, $|a;k\rangle\langle a;k|b;j\rangle$ is the projection of the ket vector $|b;j\rangle$ along the vector $|a;k\rangle$. Later $|a;k\rangle$ will be redefined in terms of a phase factor, that is, a constant multiplier of norm 1, but then $|a;k\rangle\langle a;k|$ is independent of such a phase factor. These projectors are the starting point for defining the operator connected to the e-variable λ^a:

$$A^a = \sum_k u_k |a;k\rangle\langle a;k|. \tag{7}$$

Since λ^a was assumed to be maximal, all the values u_k must be different. Thus A^a is an operator with distinct eigenvalues. All the eigenvalues and eigenvectors can be recovered by specifying the operator A^a. Since the eigenvalues are real-valued, A^a is per definition Hermitian: $A^{a\dagger} = A^a$ (see Appendix 3).

Interpretation of the operator A^a: *This gives all possible states and all possible values corresponding to the maximal accessible e-variable λ^a.*

13 The general symmetrical epistemic setting

Go back to the definition of the maximal symmetrical epistemic setting. Let again ϕ be the inaccessible conceptual variable and let λ^a for $a \in \mathcal{A}$ be the maximal accessible conceptual variables satisfying Assumption 1. Let the corresponding induced groups G^a and G satisfy Assumption 2. Finally, let t^a for each a be an arbitrary function on the range of λ^a, and assume that we observe $\theta^a = t^a(\lambda^a); a \in \mathcal{A}$. We will call this the symmetrical epistemic setting; it is no longer necessarily maximal with respect to the observations θ^a.

Consider first the quantum states $|a;k\rangle$. We are no longer interested in the full information on λ^a, but keep the Hilbert space as in Section 12, and now let $f^a_k(\phi) = I(t^a(\lambda^a) = t^a(u_k)) = I(\theta^a = u^a_k)$, where $u^a_k = t^a(u_k)$. We let again g'_{0a1} and g'_{0a2} be two distinct elements of G' such that $g'_{0ai} \to g_{0a}$, define $g'_a = (g'_{0a1})^{-1} g'_{0a2}$ and then

$$|a;k\rangle = V(g'_a)U_a f^a_k = V(g'_a)|0;k\rangle,$$

where $|0;k\rangle = f^0_k$.

INTERPRETATION OF THE STATE VECTOR $|a;k\rangle$: *1) The question: 'What is the value of θ^a?' has been posed. 2) We have obtained the answer $\theta^a = u^a_k$. Both the question and the answer are contained in the state vector.*

From this we may define the operator connected to the e-variable θ^a:

$$A^a = \sum_k u^a_k |a;k\rangle\langle a;k| = \sum_k t^a(u_k)|a;k\rangle\langle a;k|.$$

Then A^a is no longer necessarily an operator with distinct eigenvalues, but A^a is still Hermitian: $A^{a\dagger} = A^a$.

INTERPRETATION OF THE OPERATOR A^a: *This gives all possible states and all possible values corresponding to the accessible e-variable θ^a.*

The projectors $|a;k\rangle\langle a;k|$ and hence the ket vectors $|a;k\rangle$ are no longer uniquely determined by A^a: They can be transformed arbitrarily by unitary transformations in each space corresponding to one eigenvalue. In general I will redefine $|a;k\rangle$ by allowing it to be subject to such transformations. These transformed eigenvectors all still correspond to the same eigenvalue, that is, the same observed value of θ^a and they give the same operators A^a. In particular, in the maximal symmetric epistemic setting I will allow an arbitrary constant phase factor in the definition of the $|a;k\rangle$'s.

As an example of the general construction, assume that λ^a is a vector: $\lambda^a = (\theta^{a_1}, ..., \theta^{a_m})$. Then one can write a state vector corresponding to λ^a as

$$|a;k\rangle = |a_1;k_1\rangle \otimes ... \otimes |a_m;k_m\rangle$$

in an obvious notation, where $a = (a_1, ..., a_m)$ and $k = (k_1, ..., k_m)$. The different θ's may be connected to different subsystems.

So far I have kept the same groups G^a and G when going from λ^a to $\theta^a = t^a(\lambda^a)$, that is from the maximal symmetrical epistemic setting to the general symmetrical epistemic setting. This implies that the (large) Hilbert space will be the same. A special case occurs if t^a is a reduction to an orbit of G^a. This is the kind of model reduction discussed in Section 3. Then the construction of the previous sections can also be carried with a smaller group action acting just upon an orbit, resulting then in a smaller Hilbert space. In the example of the previous paragraph it may be relevant to consider one Hilbert space for each subsystem. The large Hilbert space is however the correct space to use when the whole system is considered.

Connected to a general physical system, one may have many e-variables θ and corresponding operators A. In the ordinary quantum formalism, reviewed in the next section, there is well-known theorem saying that, in my formulation, $\theta^1, ..., \theta^n$ are compatible, that is, there exists an e-variable λ such that $\theta^i = t^i(\lambda)$ for some functions t^i if and only if the corresponding operators commute:

$$[A^i, A^j] \equiv A^i A^j - A^j A^i = 0 \text{ for all } i, j.$$

(See Holevo, 2001.) Compatible e-variables may in principle be estimated simultaneously with arbitrary accuracy.

14 The quantum-mechanical culture

In this section I will no longer assume the symmetrical epistemic setting, and we thus will dispense with group-theoretical assumptions like Assumption 1 and Assumption 2. I just take as a point of departure a finitedimensional complex vector space H with kets $|k\rangle$ and corresponing bras $\langle k|$. The one-dimensional predictors $|k\rangle\langle k|$ are defined as before, and all operators on H are of the form

$$A = \sum_k v_k |k\rangle\langle k| \text{ with } \langle i|j\rangle = \delta_{ij}.$$

From this all the features of elementary quantum mechanics follow except the probability statements, which I will come back to later. The operators can be multiplied as also discussed in Appendix 3. The multiplication is associative, but not commutative. As usual we define the commutator as

$$[A, B] = AB - BA.$$

The Hermitian adjoint operator $A^\dagger$ is defined such that the ket $A^\dagger|k\rangle$ corresponds to the bra $\langle k|A$, in other words $\langle i|A^\dagger j\rangle = \langle iA|j\rangle$ for all $|i\rangle$, $|j\rangle$. This $A^\dagger$ can also be defined by complex conjugating the eigenvalues in the formula above. The observables are defined as the Hermitian operators: $A^\dagger = A$. In general one has $(AB)^\dagger = B^\dagger A^\dagger$. The possible values of the observables are their eigenvalues, and the states are given by the ket vectors.

We see that all the features of the previous section occur again, only in an abstract setting.

In an attempt to make this a little more concrete, look again at Example 16. Let J_x, J_y and J_z be the operators corresponding to spin in 3 orthogonal directions x, y and z. In quantum mechanical texts (see Messiah, 1969) it is shown that there is a constant d such that these operators satisfy the commutation relations:

$$[J_x, J_y] = idJ_z, \quad [J_y, J_z] = idJ_x, \quad [J_x, J_z] = idJ_y.$$

This can also be proved fairly easily directly in my setting for the electron spin case $j = 1/2$, using the geometry of SU(2). In standard quantum mechanics $d = \hbar/2$, where $\hbar$ is Planck's constant. We will choose units such that $d = 1/2$. In great generality, commutation relations may be derived from group properties by exploiting the relation between Lie groups and Lie algebras; see for instance Barut and Raczka (1985).

Several consequences of the above commutation relations are derived in standard texts, for instance Messiah (1969). First it is shown that J_z commutes with $J^2 = J_x^2 + J_y^2 + J_z^2$ and that J^2 has eigenvalues of the form $j(j+1)$, where j is integer or half integer. It is well known (see also the previous section) that commuting operators can be simultaneously diagonalized. In terms of the corresponding e-variables $(\theta_x, \theta_y, \theta_z)$ this means that the vector $(\|\theta\|^2, \theta_z)$ is accessible, where $\|\theta\|^2 = \theta_x^2 + \theta_y^2 + \theta_z^2$. Given j, the eigenvalues of J_z are of the form $-j, -j+1, ..., j-1, j$ as anticipated in Example 16. Also, eigenvectors can be explicitly discussed.

We conclude from this that we have two possible situations:

1) J^2 is known; more explicitly, the squared modulus $\|\theta\|^2$ is known, and takes one of the values $j(j+1)$. Then the situation is exactly as in Example 16, in particular the assumptions of the maximal symmetrical epistemic setting are satisfied.

2) The squared modulus $\|\theta\|^2$ is unknown. Then the operator $J^2 \otimes J_z$ (taking infinitely many, but discrete values) can be diagonalized, can be understood in terms of conceptual variables, but is not directly given in terms of a maximal symmetrical epistemic setting.

In conclusion, the assumptions from Section 13 defining a symmetrical epistemic setting are sometimes satisfied, sometimes not for a given quantum mechanical situation, but the introduction of conceptual variables does seem to be useful for understanding what is going on. Model reduction seems to be crucial here.

Let now by a slight change of notation $\boldsymbol{J} = (J_x, J_y, J_z)$ be the inaccessible total angular momentum of a system of particles where $\|\boldsymbol{J}\|^2 = J(J+1)$ is known. Assume that $\boldsymbol{J}$ is the sum of two spins $\boldsymbol{j}_1$ and $\boldsymbol{j}_2$ where $\|\boldsymbol{j}_1\|^2 = j_1(j_1+1)$ and $\|\boldsymbol{j}_2\|^2 = j_2(j_2+1)$ are known. Let $|m_i\rangle$ be the state where $j_{iz} = m_i$ for $-j_i \leq m_i \leq j_i$. Then the state $|M\rangle$ where $J_z = M$ can be decomposed into

$$|M\rangle = \sum_{m_1 m_2} c_{M m_1 m_2} |m_1\rangle \otimes |m_2\rangle.$$

The coefficients $c_{M m_1 m_2}$, nonzero only for $m_1 + m_2 = M$, are called Clebsch-Gordon coefficients and are discussed in standard quantum mechanical texts like Messiah (1969). Generalizations, only more technically involved, exist when $\boldsymbol{J}$ is the sum of more than two spins or angular momenta. This is the second instance where new states are found by taking linear combinations of a basic set of state vectors.

From elementary quantum mechanical texts one can get the impression that all linear combinations of state vectors in a Hilbert space are possible state vectors. This is however not true; I will discuss superselection rules later. Nevertheless, taking linear combinations of state vectors leads to the introduction of interesting and important quantum mechanical phenomena, in particular that of *entanglement*, which will be treated in Section 20.

15 Continuous e-variables. Phase space

Consider the one-dimensional movement of a single non-relativistic particle in some force field, the particle having position ξ and momentum π at some given time. Both ξ and π are e-variables and can be estimated by suitable experiments. But it has been well known from the early days of quantum mechanics that it is impossible to estimate the vector $\phi = (\xi, \pi)$ with arbitrary accuracy. Thus the point ϕ in the phase space is an inaccessible e-variable.

I will first concentrate in the position ξ. This is a continuous variable, so a state cannot be defined as simply as in the discrete case. Consider a fixed

confidence interval or credibility interval $(\underline{\xi}, \bar{\xi})$ for this position. Either ξ lies in this interval or it does not lie in this interval. In the first case, the confidence coefficient/credibility coefficient of the interval can be made arbitrarily close to 1 by doing a suitable large experiment. In the second case, the same coefficient can be made arbitrarily close to 0. Thus it is crucial by experiment, that is, by an epistemic process, to make a choice between the two indicator variables:

$$I_1(\xi) = I(\xi \in (\underline{\xi}, \bar{\xi})), \quad I_2(\xi) = I(\xi \notin (\underline{\xi}, \bar{\xi})).$$

Let G^ξ be the translation group on the real line $\mathcal{R}$. The invariant measure corresponding to G^ξ is the Lebesgue measure $d\xi$, and I will define the Hilbert space $H = L^2(\mathcal{R}, d\xi)$. The indicator $I_1(\xi)$ belongs to this space. The indicator $I_2(\xi)$ does not belong to H, but this is not important since $I_2 = 1 - I_1$.

By letting $\underline{\xi}$ and $\bar{\xi}$ vary, the ξ-state of the system can be defined in terms of the indicators I_1. For fixed $\underline{\xi}$ and $\bar{\xi}$ this is a discrete e-variable taking values $I_1 = 0$ and $I_1 = 1$. It is crucial in quantum mechanics that linear combinations of states defined by indicators and the limits of these also can be introduced as states. They will emerge through the time development of states through the Schrödinger equation; see Section 23. In fact this is the third instance where new states are found by taking linear combinations of a basic set of state vectors.

The approach I will take here is a limiting operation obtained through dividing the real line into many intervals such that the width of each interval tends to zero. Through this limiting process we can approximate any function f in H. In traditional quantum mechanics, any such f is describing a state of the particle, and f is called a wave function. I will not go into any interpretation of this here, but just mention that there are interpretations trying to connect this to the theory of stochastic processes; this is the content of the stochastic mechanics of Edward Nelson (1967). I will discuss this later in connection to the Schrödinger equation, but here I only address the limiting process. I will limit myself to continuous f.

Thus for each n let $\xi_{n1} < \xi_{n2} < ... < \xi_{nk_n}$ be a sequence of real numbers such that

1. $\xi_{n1} \to -\infty$ and $\xi_{nk_n} \to \infty$ as $n \to \infty$.

2. $\delta_n = \xi_{n,i+1} - \xi_{ni}$ is constant for $i = 1, ..., k_n - 1$ and tends to 0 as $n \to \infty$.

Let $I_{ni}(\xi) = I(\xi \in (\xi_{ni}, \xi_{n,i+1}])$ for $i = 1, 2, ..., k_n - 1$. For a given function $f \in H$, define the step function approximation f_n by $f_n(\xi) = f(\xi_{ni})$ for $\xi_{ni} \leq \xi < \xi_{n,i+1}$ when $i = 1, ..., k_n - 1$; $f_n(\xi) = 0$ for $\xi < \xi_{n1}$ and for $\xi \geq \xi_{nk_n}$. Thus $f_n(\xi) = \sum_i f(\xi_{ni}) I_{ni}(\xi)$, a linear function of indicators. Finally, on the space of such step functions define the operator A_n by

$$A_n f_n(\xi) = \sum_i \xi_{ni} f(\xi_{ni}) I_{ni}(\xi). \tag{8}$$

The interpretation of (8) is as follows: Approximate ξ by ξ_{ni} when $\xi \in (\xi_{ni}, \xi_{n.i+1}]$ (and neglect its value when $\xi < \xi_{n1}$ or $\xi \geq \xi_{nk_n}$; this is assumed to

have negligible probability/ confidence coefficient). This approximate variable is discrete, so we can use the theory of Section 12 (in the simple case where we have only one Hilbert space, so $a = 0$). The indicators I_{ni} can be regarded as an orthonormal set of ket vectors for this approximate variable for a suitable normalization of the Lebesgue measure. Then (8) is equivalent to (7) of Section 12 with $|0; i\rangle = I_{ni}(\xi)$ constituting an orthonormal basis for a Hilbert space H_n of step functions, a subspace of H. Thus A_n is the quantum-mechanical operator of the discrete variable.

It is of interest to see what happens when n tends to ∞. The following basic result is proved in Appendix 4.

THEOREM 6. *Assume that f is a continuous function in H such that the function k defined by $k(\xi) = \xi f(\xi)$ satisfies $\|k\| < \infty$. Then $\|f_n - f\| \to 0$ and $\|A_n f_n - k\| \to 0$ as $n \to \infty$.*

In this specific sense the operator A corresponding to the e-variable ξ can be said to be the operator of multiplying with ξ. By Theorem 6 it is motivated as such an operator defined on all continuous f in H such that $\int |\xi f(\xi)|^2 d\xi < \infty$.

The operator A is an unbounded operator, and as such it must always have a limited domain of definition $D(A)$. There is a very large and advanced mathematical theory on unbounded operators; see for instance Murphy (1990) or Bing-Ren (1992).

So far I have considered the position ξ. A completely parallel discussion can be made on the moment π in the Hilbert space $H^\pi = L^2(\mathcal{S}, d\pi)$, where $\mathcal{S}$ is the line where π varies. Thus the operator B corresponding to momentum π in this space is multiplication by π with domain of definition $D(B) = \{f \in H^\pi : \int |\pi f(\pi)|^2 d\pi < \infty$.

As in the discrete case it is important to have everything described in one Hilbert space, so we need a unitary transformation from H^π to H. For this case we have a completely different and simpler solution than I offered in the maximal symmetric epistemic setting, namely the use of Fourier transform. If $\widehat{f} \in H^\pi$, we define the corresponding $f \in H$ by

$$(U\widehat{f})(\xi) = f(\xi) = \frac{1}{\sqrt{2 \cdot 3.14}} \int \exp(i\frac{\xi\pi}{\hbar})\widehat{f}(\pi)d\pi,$$

where $\hbar$ is Planck's constant, which has the correct unit of measurement. One point here is that this unitary transformation does not transform indicator variables into indicator variables, so there is no confusion between simple π-states and simple ξ-states. The inverse transformation is given by

$$(U^\dagger f)(\pi) = \widehat{f}(\pi) = \frac{1}{\sqrt{2 \cdot 3.14}} \int \exp(-i\frac{\xi\pi}{\hbar})f(\xi)d\xi.$$

By partial integration one can show that the operator $C = UBU^\dagger$ corresponding to B in H is given by $-i\hbar\frac{d}{d\xi}$ with domain of definition $D(C)$ given by the set of differentiable f such that $\int |f'(\xi)|^2 d\xi < \infty$. It follows that when

$f \in D(A) \cap D(C)$ we have $(AC - CA)f(\xi) = i\hbar f(\xi)$, so A and C do not commute. Hence by the brief discussion at the end of Section 13, ξ and π cannot be estimated simultaneously with arbitrary accuracy, in agreement with observed fact. From the commutation relation Heisenberg's uncertainty relation can be proved: From any estimators $\widehat{\xi}$ and $\widehat{\pi}$ we have $\mathrm{std}(\widehat{\xi})\mathrm{std}(\widehat{\pi}) > \hbar/2$. For a derivation, see standard quantum-mechanical texts or Holevo (2001). But now I am anticipating the inference theory which will be developed in the following sections.

16 Link to statistical inference

In this section I again assume first the maximal symmetrical epistemic setting of Section 11, but everything that I say can be generalized, see later. We can here think of a spin component in a fixed direction to be assessed. To assume a state $|a;k\rangle$ is to assume perfect knowledge of the e-variable λ^a: $\lambda^a = u_k$. Such perfect knowledge is rarely available. In practice we have data z^a about the system, and use these data to obtain knowledge about λ^a. Let us start with Bayesian inference. This assumes prior probabilities π_k^a on the values u_k, and after the inference we have posterior probabilities $\pi_k^a(z^a)$. In either case we summarize this information in the density operator:

$$\sigma^a = \sum_k \pi_k^a |a;k\rangle\langle a;k|.$$

Interpretation of the density operator σ^a: *1) We have posed the question 'What is the value of λ^a?' 2) We have specified a prior or posterior probability distribution π_k^a over the possible answers. The probability for all possible answers to the question, formulated in terms of state vectors, can be recovered from the density operator.*

A third possibility for the probability specifications is a confidence distribution; see Subsections 2.3 and 2.4 and references there. For discrete λ^a the confidence distribution function H^a is connected to a discrete distribution, which gives the probabilities π_k^a. Extending the argument of Xie and Singh (2011) to this situation, this should not be looked upon as a distribution *of* λ^a, but a distribution *for* λ^a, to be used in the epistemic process.

Since the sum of the probabilities is 1, the trace (sum of eigenvalues) of any density operator is 1. In the quantum mechanical literature, a density operator is any positive operator with trace 1.

Note that specification of the maximally accessible e-variables λ^a is equivalent to specifying $t(\lambda^a)$ for any one-to-one function t. The operator $t(A^a)$ has then distinct eigenvalues since the operator A^a has distinct eigenvalues. Hence it is enough in order to specify the question 1) to give the set of orthonormal vectors $|a;k\rangle$.

Given the question a, the e-variable λ^a plays the role similar to a parameter in statistical inference, even though it may be connected to a single unit.

Inference can be done by preparing many independent units in the same state. Inference is then from data z^a, a part of the total data z that nature can provide us with. All inference theory that one finds in standard texts like Lehmann and Casella (1998) applies. In particular, the concepts of unbiasedness, equivariance, avarage risk optimality, minimaxity and admissibility apply. None of these concepts are much discussed in the physical literature, first because measurements there are often considered as perfect, at least in elementary texts, secondly because, when measurements are considered in the physical literature, they are discussed in terms of the more abstract concept of an operator-valued measure, which is relevant if the question a is not kept fixed; see later.

Whatever kind of inference we make on λ^a, we can take as a point of departure the statistical model and the generalized likelihood principle of Section 8. Hence after an experiment is done, and given some context τ, all evidence on λ^a is contained in the likelihood $p(z^a|\tau, \lambda^a)$, where z^a is the portion of the data relevant for inference on λ^a, also assumed discrete. This is summarized in the likelihood effect:

$$E(z^a, \tau) = \sum_k p(z^a|\tau, \lambda^a = u_k)|a; k\rangle\langle a; k|.$$

Interpretation of the likelihood effect $E(z^a, \tau)$: *1) We have posed some inference question on the maximally accessible parameter λ^a. 2) We have specified the relevant likelihood for the data. The likelihood for all possible answers of the question, formulated in terms of state vectors, can be recovered from the likelihood effect.*

Since the focused question assumes discrete data, each likelihood is in the range $0 \leq p \leq 1$. In the quantum mechanical literature, an effect is any operator with eigenvalues in the range $[0, 1]$.

Return now to the generalized likelihood principle of Section 8. The following principle follows.

The focused generalized likelihood principle (FGLP) *Consider two potential experiments in the symmetrical epistemic setting with equivalent contexts τ, and assume that the inaccessible conceptual variable ϕ is the same in both experiments. Suppose that the observations z_1^* and z_2^* have proportional likelihood effects in the two experiments, with a constant of proportionality independent of the conceptual variable.*

Assume in addition that one is sure that the decision problem does not depend on any irrelevant UCV. Then the questions posed in the two experiments are equivalent, that is , there is a maximal e-variable λ^a which can be considered to be the same in the two experiments, and the two observations produce the same evidence on λ^a in this context.

In many examples the two observations will have equal, not only proportional, likelihood effects. Proportionality of the likelihood may be an option

when the e-variable is not maximal. Here is an example: Assume $p(z^a|\tau,\theta^a) = p(-z^a|\tau,\theta^a) = p(z^a|\tau,-\theta^a) = p(-z^a|\tau,-\theta^a)$. Then $z^b = (z^a)^2$ contains the same evidence on θ^a as z^a, we have only evidence on $\theta^b = (\theta^a)^2$, we have $p(z^b|\tau,\theta^a) = 2p(z^a|\tau,\theta^a)$ and $|b;k^2\rangle\langle b;k^2| = |a;k\rangle\langle a;k| + |a;-k\rangle\langle a;-k|$, where $|b;k^2\rangle$ corresponds to the question: 'What is the value of θ^b?' with answer $\theta^b = (u_k^a)^2$, and $|a;-k\rangle$ corresponds to the question: 'What is the value of θ^a?' with the answer $\theta^a = -u_k^a$. A similar situation occurs whenever z^a and θ^a are multi-valued in a corresponding way. In the following discussion I will specialize to the case of one experiment and equal likelihood. Then the FGLP says simply that the experimental evidence is a function of the likelihood effect.

In the FGLP we have the freedom to redefine the e-variable in the case of coinciding eigenvalues in the likelihood effect, that is, if $p(z^a|\tau,\lambda^a = u_k) = p(z^a|\tau,\lambda^a = u_l)$ for some k, l. An extreme case is the likelihood effect $E(z^a,\tau) = I$, where all the likelihoods are 1, that is, the probability of z is 1 under any considered model. Then any maximal accessible e-variable λ^a will serve our purpose.

17 Rationality and experimental evidence

This section may at first sight seem to be slightly more speculative than the rest of the paper, but it will end with a very concrete result.

Throughout the section I will consider a fixed context τ and a fixed maximal epistemic setting in this context. The inaccessible e-variable is ϕ, and I assume that the maximal accessible e-variables λ^a take a discrete set of values. Let the data behind the potential experiments be z^a, also assumed to take a discrete set of values.

Let first a single experimentalist A be in this situation, and let all conceptual variables be attached to A, although he also has the possibility to receiving information from others through part of the context τ. He has the choice of doing different experiments a, and he also has the choice of choosing different models for his experiment through his likelihood $p(z^a|\tau,\lambda^a)$. The experiment and the model, hence the likelihood, should be chosen before the data are obtained. All these choices are summarized in the likelihood effect E, a function of the at present unknown data z^a, and also of the unknown e-variable λ^a. For use after the experiment, he should also choose a good estimator/predictor $\widehat{\lambda^a}$, and he may also have to choose some loss function, but the principles behind these latter choices will be considered as part of the context τ. If he chooses to do a Bayesian analysis, the estimator should be based on a prior $\pi(\lambda^a|\tau)$. We assume that A is trying to be as rational as possible in all his choices, and that this rationality is connected to his loss function or to other criteria.

What should be meant by experimental evidence, and how should it be measured? As a natural choice, let the experimental evidence that we are seeking, be the marginal probability of the obtained data for a fixed experiment and for a given likelihood function. From the experimentalist A's point of view this is

given by:

$$p_A^a(z^a|\tau) = \sum_k p_A(z^a|\tau, \lambda^a = u_k)\pi_A(\lambda^a = u_k|\tau),$$

assuming the likelihood chosen by A and A's prior π_A for λ^a. Some Bayesians claim that their own philosophy is the only one which is consistent with the likelihood principle. For my own view on this, see below and also comments in Section 8. In a non-Bayesian analysis, we can let $p_A^a(z^a|\tau)$ be the probability given the true value u_k^0 of the parameter: $p_A^a(z^a|\tau) = p_A(z^a|\tau, \lambda^a = u_k^0)$. In general, take $p_A^a(z^a|\tau)$ as the probability of the part of the data z^a which A assesses in connection to his inference on λ^a. By the FGLP - specialized to the case of one experiment and equal likelihoods - this experimental evidence must be a function of the likelihood effect: $p_A^a(z^a|\tau) = q_A(E(z^a)|\tau)$.

We have to make precise in some way what is meant by the rationality of the experimentalist A. He has to make many difficult choices on the basis of uncertain knowledge. His actions can partly be based on intuition, partly on experience from similar situations, partly on a common scientific culture and partly on advices from other persons. These other persons will in turn have their intuition, their experience and their scientific education. Often A will have certain explicitly formulated principles on which to base his decisions, but sometimes he has to dispense with the principles. In the latter case, he has to rely on some 'inner voice', a conviction which tells him what to do.

We will formalize all this by introducing a perfectly rational superior actor D, to which all these principles, experiences and convictions can be related. We also assume that D can observe everything that is going on, in particular A, and that he on this background can have some influence on A's decisions. The real experimental evidence will then be defined as *the probability of the data z^a from D's point of view, which we assume also to give the real objective probabilities.* By the FGLP this must again be a function of the likelihood effect E, where the likelihood now may be seen as the objectively correct model.

$$p^a(z^a|\tau) = q(E(z^a)|\tau) \tag{9}$$

As said, we assume that D is perfectly rational. This can be formalized mathematically by considering a hypothetical betting situation for D against a bookie, nature N. A similar discussion was recently done using a more abstract language by Hammond (2011). Note the difference to the ordinary Bayesian assumption, where A himself is assumed to be perfectly rational. This difference is crucial to me. I do not see any human scientist, including myself, as being perfectly rational. We can try to be as rational as possible, but we have to rely on some underlying rational principles that partly determine our actions.

So let the hypothetical odds of a given bet for D be $(1-q)/q$ to 1, where q is the probability as defined by (9). This odds specification is a way to make precise that, given the context τ and given the question a, the bettor's probability that the experimental result takes some value is given by q: For a given utility measured by x, the bettor D pays in an amount qx - the stake - to the bookie. After the experiment the bookie pays out an amount x - the payoff - to the

bettor if the result of the experiment takes the value z^a, otherwise nothing is payed.

The rationality of D is formulated in terms of

THE DUTCH BOOK PRINCIPLE. *No choice of payoffs in a series of bets shall lead to a sure loss for the bettor.*

For a related use of the same principle, see Caves et al. (2002).

ASSUMPTION 3. *Consider in some context τ a maximal symmetrical epistemic setting where the FGLP is satisfied, and the whole situation is observed and acted upon by a superior actor D as described above. Assume that D's probabilities q given by (9) are taken as the experimental evidence, and that D acts rationally in agreement with the Dutch book principle.*

A situation where all the three assumptions 1, 2 and 3 hold will be called a *rational epistemic setting*. It will be assumed to be implied by essential situations of quantum mechanics. Later the assumptions 1 and 2 will be weakened. The question will also later be raised if it can be coupled to certain macroscopic situations.

THEOREM 7. *Assume a rational epistemic setting. Let E_1 and E_2 be two likelihood effects in this setting, and assume that $E_1 + E_2$ also is a likelihood effect. Then the experimental evidences, taken as the probabilities of the corresponding data, satisfy*

$$q(E_1 + E_2|\tau) = q(E_1|\tau) + q(E_2|\tau).$$

Proof.

The result of the theorem is obvious, without making Assumption 3, if E_1 and E_2 are likelihood effects connected to experiments on the same e-variable λ^a. We will prove it in general. Consider then any finite number of potential experiments including the two with likelihood effects E_1 and E_2. Let $q_1 = q(E_1|\tau)$ be equal to (9), and let $q_2 = q(E_2|\tau)$ be equal to the same quantity with a replaced by b. Consider in addition the following randomized experiment: Throw an unbiased coin. If head, choose the experiment with likelihood effect E_1; if tail, choose the experiment with likelihood effect E_2. This is a valid experiment. The likelihood effect when the coin shows head is $\frac{1}{2}E_1$, when it shows tail $\frac{1}{2}E_2$, so that the likelihood effect of this experiment is $E_0 = \frac{1}{2}(E_1 + E_2)$. Define $q_0 = q(E_0)$. Let the bettor bet on the results of all these 3 experiments: Payoff x_1 for experiment 1, payoff x_2 for experiment 2 and payoff x_0 for experiment 0.

I will divide into 3 possible outcomes: Either the likelihood effect from the data z is E_1 or it is E_2 or it is none of these. The randomization in the choice of E_0 is considered separately from the result of the bet. (Technically this can be done by repeating the whole series of experiments many times with the same

randomization. This is also consistent with the conditionality principle.) Thus if E_1 occurs, the payoff for experiment 0 is replaced by the expected payoff $x_0/2$, similarly if E_2 occurs. The net expected amount the bettor receives is then

$$x_1 + \frac{1}{2}x_0 - q_1x_1 - q_2x_2 - q_0x_0 = (1-q_1)x_1 - q_2x_2 - (1-2q_0)\frac{1}{2}x_0 \text{ if } E_1,$$

$$x_2 + \frac{1}{2}x_0 - q_1x_1 - q_2x_2 - q_0x_0 = -q_1x_1 - (1-q_2)x_2 - (1-2q_0)\frac{1}{2}x_0 \text{ if } E_2,$$

$$-q_1x_1 - q_2x_2 - 2q_0 \cdot \frac{1}{2}x_0 \text{ otherwise.}$$

The payoffs (x_1, x_2, x_0) can be chosen by nature N in such a way that it leads to sure loss for the bettor D if not the determinant of this system is zero:

$$0 = \begin{vmatrix} 1-q_1 & -q_2 & 1-2q_0 \\ -q_1 & 1-q_2 & 1-2q_0 \\ -q_1 & -q_2 & -2q_0 \end{vmatrix} = q_1 + q_2 - 2q_0.$$

Thus we must have

$$q(\frac{1}{2}(E_1 + E_2)|\tau) = \frac{1}{2}(q(E_1|\tau) + q(E_2|\tau)).$$

If $E_1 + E_2$ is an effect, the common factor $\frac{1}{2}$ can be removed by changing the likelihoods, and the result follows.

COROLLARY. *Assume a rational epistemic setting. Let E_1, E_2, ... be likelihood effects in this setting, and assume that $E_1 + E_2 + ...$ also is a likelihood effect. Then*

$$q(E_1 + E_2 + ...|\tau) = q(E_1|\tau) + q(E_2|\tau) + ...$$

Proof.
The finite case follows immediately from Theorem 7. Then the infinite case follows from monotone convergence.

The result of this section is quite general. In particular the loss function and any other criterion for the success of the experiments are arbitrary. So far I have assumed that the choice of experiment a is fixed, which implies that it is the same for A and for D. However, the result also applies to the following more general situation: Let A have some definite purpose of his experiment, and to achieve that purpose, he has to choose the question a in a clever manner, as rationally as he can. Assume that this rationality is formalized through the actor D, who has the ideal likelihood effect E and the experimental evidence $p(z|\tau) = q(E|\tau)$. If two such questions shall be chosen, the result of Theorem 7 holds, with essentially the same proof.

17.1 On the nature of the superior actor D

We all go through our lives making repeated choices in different contexts. These choices are governed by our free will, but they may also be influenced by people that we look up to, who perhaps have done similar choices before. In our childhood, the persons that form our basis are most often our parents, but later other ideals may take over. Human beings that get a confused relation to their first ideals, may later have difficulties in making good choices, and in certain cases they may end up with suffering from serious psychological defects, even mental illnesses.

As scientists we also have ideals that we look up to. These may be personal, or they may be substantiated through certain well-defined principles. Earlier in this section I made the assumption that the experimentalist A, when posing a focused question to nature, made his decisions inspired by an ideal D, and that D was perfectly rational. This may be regarded as a simplification. In reality, when making our choices, we are influenced by a multitude of conscious or subconscious causes. All these causes are here collected together in the actor D. I assume that D has a positive influence on A, positive with respect to the question that A has chosen as the focus of his experiment.

Let us look at the process of making choices in some greater generality. People in different cultures make their choices partly on the basis of cultural values. These values may have a historical origin, and they may also be related to religion. Christianity, Islam and Judaism are all founded upon the belief in a personal God. This belief is intimately connected to different, unfortunately partly conflicting, cultures. The believers act under the assumption that there is a God behind everything, and that God is perfectly rational. They believe that He influences all human beings, also those who serve as ideals for others. In this sense, God may take the rôle as the ultimate ideal D within the relevant culture.

This situation is obviously not satisfactory from a scientific point of view. A human being should be free to believe in a personal God, in fact such a belief may have very positive effect on his life. But if a scientist should take such a belief as a basis for his intuitive choices, God should act over and above all human cultures, and in particular He should be independent of the way He is worshipped in any specific congregation. Nevertheless I personally see the respect for sacred values and the worshipping of God as something of the deepest and most valuable in human life.

In general a culture may be looked upon as part of a man's context when making his choices. At the outset, all human beings should be respected, and so also the context they have for making their choices. Hence it is a part of my philosophy that no culture should in principle be seen as superior to other cultures when it comes to inspiring people's choices. However, this tolerance has it limits; one of these is an ultimate respect for people's life. Extremists taking lives under the belief that their own culture is threatened by other cultures, should not in any way be accepted. In addition there are of course other universal ethical rules that should be respected.

In essence certain cultural values and more generally certain value-contexts for making choices may be seen from a global point of view to be more satisfactory than other set of values, but this can only be determined by rational arguments. Hence communication between cultures is very important in our world as it is now. As a particular continuation of this statement, this book in itself is written with the purpose of finding a common language with which one can communicate across scientific epistemic cultures.

Some people may react against me in that I am discussing aspects of religion and of culture here in relation to the motivation for a purely mathematical result, in fact a relatively simple result. But this is a result which in the next section will form the basis for deriving a formula by which one can calculate probabilities in quantum mechanics. I will later, in Section 22, come back to the insight that our free will can be mimicked by nature, and this is connected with deep aspects of quantum mechanics itself.

18 The Born formula

18.1 The basic formula

Born's formula is the basis for all probability calculations in quantum mechanics. In textbooks it is usually stated as a separate axiom, but it has also been argued for by using various sets of assumptions; see Helland (2008) for some references. Here I will base the discussion upon the result of Section 17.

I begin with a recent result by Busch (2003), giving a new version of a classical mathematical theorem by Gleason. Busch's version has the advantage that it is valid for a Hilbert space of dimension 2, which Gleason's original theorem is not, and it also has a simpler proof. For completeness I reproduce the proof for the finite-dimensional case in Appendix 5.

Let in general H be any Hilbert space. Recall that an effect E is any operator on the Hilbert space with eigenvalues in the range $[0,1]$. A generalized probability measure μ is a function on the effects with the properties

(1) $0 \leq \mu(E) \leq 1$ for all E,
(2) $\mu(I) = 1$,
(3) $\mu(E_1 + E_2 + ...) = \mu(E_1) + \mu(E_2) + ...$ whenever $E_1 + E_2 + ... \leq I$.

THEOREM 8. (BUSCH, 2003). *Any generalized probability measure μ is of the form $\mu(E) = \text{trace}(\sigma E)$ for some density operator σ.*

It is now easy to see that $q(E|\tau) = p(z|\tau)$ on the ideal likelihood effects of Section 17 is a generalized probability measure if Assumption 3 holds: (1) follows since q is a probability; (2) since $E = I$ implies that the likelihood is 1 for all values of the e-variable, hence $p(z) = 1$; finally (3) is a concequence of the corollary of Theorem 7. Hence there is a density operator $\sigma = \sigma(\tau)$ such that $p(z|\tau) = \text{trace}(\sigma(\tau)E)$ for all ideal likelihood effects $E = E(z)$.

Define now a *perfect experiment* as one where the measurement uncertainty can be disregarded. The quantum mechanical literature operates very much with perfect experiments which give well-defined states $|k\rangle$. From the point of view of statistics, if, say the 99% confidence or credibility region of λ^b is the single point u_k, we can infer approximately that a perfect experiment has given the result $\lambda^b = u_k$.

In our maximal symmetric epistemic setting then: We have asked the question: What is the value of the maximally accessible e-variable λ^b, and are interested in finding the probability of the answer $\lambda^b = u_j$ though a perfect experiment. This is the probability of the state $|b; j\rangle$. Assume now that this probability is sought in a context $\tau = \tau^{a,k}$ defined as follows: We have previous knowledge of the answer $\lambda^a = u_k$ of another maximal question: What is the value of λ^a? That is, we know the state $|a; k\rangle$. If λ^a is maximally accessible, this is the maximal knowledge about the system that τ may contain, so the context τ cannot contain more information about this system. It can contain irrelevant information, however.

THEOREM 9. (BORN'S FORMULA) *Assume a rational epistemic setting. In the above situation we have:*

$$P(\lambda^b = u_j | \lambda^a = u_k) = |\langle a; k | b; j \rangle|^2.$$

Proof.

Fix j and k, let $|v\rangle$ be either $|a; k\rangle$ or $|b; j\rangle$, and consider likelihood effects of the form $E = |v\rangle\langle v|$. This corresponds in both cases to a perfect measurement of a maximally accessible parameter with a definite result. By Theorem 8 there exists a density operator $\sigma^{a,k} = \sum_i \pi_i(\tau^{a,k})|i\rangle\langle i|$ such that $q(E|\tau^{a,k}) = \langle v|\sigma^{a,k}|v\rangle$, where $\pi_i(\tau^{a,k})$ are non-negative constants adding to 1. Consider first $|v\rangle = |a; k\rangle$. For this case one must have $\sum_i \pi_i(\tau^{a,k})|\langle i|a; k\rangle|^2 = 1$ and thus $\sum_i \pi_i(\tau^{a,k})(1 - |\langle i|a; k\rangle|^2) = 0$. This implies for each i that either $\pi_i(\tau^{a,k}) = 0$ or $|\langle i|a; k\rangle| = 1$. Since the last condition implies $|i\rangle = |a; k\rangle$ (modulus an irrelevant phase factor), and this is a condition which can only be true for one i, it follows that $\pi_i(\tau^{a,k}) = 0$ for all other i than this one, and that $\pi_i(\tau^{a,k}) = 1$ for this particular i. Summarizing this, we get $\sigma^{a,k} = |ak\rangle\langle a; k|$, and setting $|v\rangle = |b; j\rangle$, Born's formula follows, since $q(E|\tau^{a,k})$ in this case is equal to the probability of the perfect result $\lambda^b = u_j$.

18.2 Consequences

Here are three easy consequences of Born's formula:

(1) If the context of the system is given by the state $|a; k\rangle$, and A^b is the operator corresponding to the e-variable λ^b, then the expected value of a perfect measurement of λ^b is $\langle a; k|A^b|a; k\rangle$.

(2) If the context is given by a density operator σ, and A is the operator corresponding to the e-variable λ, then the expected value of a perfect measurement of λ is $\text{trace}(\sigma A)$.

(3) In the same situation the expected value of a perfect measurement of $f(\lambda)$ is $\text{trace}(\sigma f(A))$.

Proof of (1):

$$\mathrm{E}(\lambda^b|\lambda^a = u_k) = \sum_i u_i P(\lambda^b = u_i|\lambda^a = u_k)$$

$$= \sum_i u_i \langle a; k|b; i\rangle\langle b; i|a; k\rangle = \langle a; k|A^b|a; k\rangle.$$

These results give an extended interpretation of the operator A compared to what I gave in Section 12: There is a simple formula for all expectations in terms of the operator. On the other hand, the set of such expectations determine the state of the system. Also on the other hand: If A is specialized to an indicator function, we get back Born's formula, so the consequences are equivalent to this formula.

As an application of Born's formula, we give the transition probabilities for electron spin. Throughout this paper, we will, for a given direction a, define the e-variable λ^a as $+1$ if the measured spin component by a perfect measurement for the electron is $+\hbar/2$ in this direction, $\lambda^a = -1$ if the component is $-\hbar/2$. Assume that a and b are two directions in which the spin component can be measured.

Proposition 5. *For electron spin we have*

$$P(\lambda^b = \pm 1|\lambda^a = +1) = \frac{1}{2}(1 \pm \cos(a \cdot b)).$$

This is proved in several textbooks, for instance Holovo (2001), from Born's formula. A proof using the Pauli spin matrices is also given in Helland (2010).

18.3 A macroscopic example

A very relevant question is now: Are all these results, including Born's formula, by necessity confined to the microworld? Recently, physicists have become interested in larger systems where quantum mechanics is valid, see Vedral (2011). As we have defined it, there is nothing microscopic about the symmetrical epistemic setting. It may or may not be that the rationality Assumption 3 also is valid for some larger scale systems. The following example illustrates the point.

Example 17. In a medical experiment, let μ_a, μ_b, μ_c and μ_d be continuous inaccessible parameters, the hypothetical effects of treatment a, b, c and d, respectively. Assume that the focus of the experiment is to compare treatment b with the mean effect of the other treatments, which is supposed to give the

parameter $\frac{1}{3}(\mu_a+\mu_c+\mu_d)$. One wants to do a pairwise experiment, but it turns out that the maximal parameter which can be estimated, is

$$\lambda^b = \text{sign}(\mu^b - \frac{1}{3}(\mu_a + \mu_c + \mu_d)).$$

(Imagine for example that one has four different ointments against rash. A patient is treated with ointment b on one side of his back; a mixture of the other ointments on the other side of his back. It is only possible to observe which side improves best, but this observation is assumed to be very accurate. One can in principle do the experiment on several patients, and select out the patients where the difference is clear.) This experiment is done on a selected set of experimental units, on whom it is known from earlier accurate experiments that the corresponding parameter

$$\lambda^a = \text{sign}(\mu^a - \frac{1}{3}(\mu_b + \mu_c + \mu_d))$$

takes the value +1. In other words, one is interested in the probabilities

$$\pi = P(\lambda^b = +1 | \lambda^a = +1).$$

Consider first a Bayesian approach. Natural priors for $\mu_a, ..., \mu_d$ are independent $N(\nu, \sigma^2)$ with the same ν and σ. By location and scale invariance, there is no loss in generality by assuming $\nu = 0$ and $\sigma = 1$. Then the joint prior of $\zeta^a = \mu_a - \frac{1}{3}(\mu_b + \mu_c + \mu_d)$ and $\zeta^b = \mu_b - \frac{1}{3}(\mu_a + \mu_c + \mu_d)$ is multinormal with mean $\mathbf{0}$ and covariance matrix

$$\begin{pmatrix} \frac{4}{3} & -\frac{4}{9} \\ -\frac{4}{9} & \frac{4}{3} \end{pmatrix}.$$

A numerical calculation from this gives

$$\pi = P(\zeta^b > 0 | \zeta^a > 0) \approx 0.43.$$

This result can also be assumed to be valid when $\sigma \to \infty$, a case which in some sense can be considered as independent objective priors for $\mu_a, ..., \mu_d$.

Now consider a rational epistemic setting for this experiment. Since again scale is irrelevant, a natural group on $\mu_a, ..., \mu_d$ is a 4-dimensional rotation group around a point $(\nu, ..., \nu)$ together with a translation of ν. Furthermore, ζ^a and ζ^b are contrasts, that is, linear combinations with coefficients adding to 0. The space of such contrasts is a 3-dimensional subspace of the original 4-dimensional space, and by a single orthogonal transformation, the relevant subset of the 4-dimensional rotations can be transformed into the group G of 3-dimensional rotations on this latter space, and the translation in ν is irrelevant. One such orthogonal transformation is given by

$$\psi_0 = \frac{1}{2}(\mu_a + \mu_b + \mu_c + \mu_d),$$

$$\psi_1 = \frac{1}{2}(-\mu_a - \mu_b + \mu_c + \mu_d),$$

$$\psi_2 = \frac{1}{2}(-\mu_a + \mu_b - \mu_c + \mu_d),$$

$$\psi_3 = \frac{1}{2}(-\mu_a + \mu_b + \mu_c - \mu_d).$$

Let G be the group of rotations orthogonal to ψ_0. We find

$$\zeta^a = -\frac{2}{3}(\psi_1 + \psi_2 + \psi_3),$$

$$\zeta^b = -\frac{2}{3}(\psi_1 - \psi_2 - \psi_3).$$

The rotation group element transforming ζ^a into ζ^b is homomorphic under G to the rotation group element g_{ab} transforming $a = -\frac{1}{\sqrt{3}}(1,1,1)$ into $b = -\frac{1}{\sqrt{3}}(1,-1,-1)$. Let G^a be the maximal subgroup of G under which ζ^a is permissible. This is isomorphic with the group of rotations around a together with a reflection in the plane perpendicular to a, but the action on ζ^a is just a reflection. The orbits of this group are given by two-point sets $\{\pm c\}$. In conclusion, the whole situation is completely equivalent to the spin-example of Example 16 and satisfies the assumptions of the symmetrical epistemic setting. Making the rationality Assumption 3 then implies from Proposition 5:

$$\pi = P(\text{sign}(\zeta^b) = +1 | \text{sign}(\zeta^{\text{a}}) = +1) = \frac{1}{2}(1 + a \cdot b) = \frac{1}{3}.$$

To be precise, Example 17 satisfies the assumptions of the symmetrical epistemic setting except Assumptions 2a) and 2c). To have these assumptions satisfied, we must extend the situation:

EXAMPLE 18. Let the situation be as in Example 17 with the addition that we have available treatments with hypothetical effects μ_a for $a \in \mathcal{A}$, where the index set $\mathcal{A}$ can be taken to be the 3-dimensional unit sphere.

It is clear that the extension from Example 17 to Example 18 does not mean anything for the result.

I guess that many statisticians will prefer the Bayesian calculations here for the rational epistemic setting calculations, which some may consider to have a more speculative foundation. But the prior chosen in this example must be considered somewhat arbitrary, and its 'objective' limit may lead to conceptual difficulties. Since experiments of this kind can in principle be done in practice - at least approximately, the question whether the Bayesian solution or the rational epistemic setting solution holds in such cases, must ultimately be seen as an empirical question. I hope to discuss further whether the rational epistemic setting can apply to certain macroscopic situations in a future publication.

18.4 Generalizations of Born's formula

Consider first the general symmetric epistemic setting of Section 13, that is, our question concerns a function $\theta^a = t^a(\lambda^a)$ of the maximal accessible e-variable.

Going through the proof of Born's formula, it is nowhere used that the accessible epistemic e-variable is maximal. At the end of Section 16, an example was shown where proportional likelihoods were used for non-maximal e-variables. However, in the proof of Theorem 7 we had a situation where equality of the likelihoods was *assumed*, so that the experimental evidence was a function of the likelihood effect. Hence when $|a;k\rangle$ and $|b;j\rangle$ refer to any symmetrical accessible e-variables θ^a and θ^b, we still have $P(|b;j\rangle||a;k\rangle) = |\langle a;k|b;j\rangle|^2$.

Born's formula can be extended beyond the symmetrical epistemic setting. First the context τ may contain irrelevant information. But also the target state $|K,b;k\rangle$ may contain irrelevant information in addition to the answer to the question 'What is the value of λ^b?': $|K,b;j\rangle = |b;j\rangle \otimes |K\rangle$. The additional information should not change the context even if it is unknown to the experimentalist; here we may appeal to case 2) of the generalized principle for conditioning (GPC) of Section 7. Also, in the spin example of Section 13 the case with unknown total spin represents no problem: $\langle i|j\rangle = 0$ when $|i\rangle$ and $|j\rangle$ have different total spin, so no transition between these states can occur. Finally, from the discussion around Example 18, it seems like the Assumptions 2a) and 2c) of Section 11 can be relaxed. My conjecture is that the discussions of this paper can be generalized to all cases of interest in quantum mechanics, but this represents in its generality an open question.

But let $|\psi_k\rangle$ and $|\theta_j\rangle$ be any ket vectors, formed by linear combinations of basic vectors or in other ways, only connected to e-variables ψ and θ. Then we can again go through our arguments for the Born formula, and see that these arguments carry over, so $P(|\theta_j\rangle||\psi_k\rangle) = |\langle\psi_k|\theta_j\rangle|^2$.

Finally, to indicate how Born's formula generalizes to continuous systems, look at the spacial wave function f of Section 15. By using a limiting argument similar to that used in Theorem 6, and anticipating the discussion of Section 23, one can prove the following from the Born formula: Assume that the state of the system is given by the wave function $f(\xi)$. Then the probability density of a perfect measurement of the position ξ is given by $|f(\xi)|^2$.

18.5 Superselection rules

Two states $|\psi\rangle$ and $|\theta\rangle$ obay a superselection rule if $\langle\psi|A|\theta\rangle = 0$ for all operators A representing physical observables. This can be the case for instance if the Hilbert space decomposes as $H = H_1 \oplus H_2$, $|\psi\rangle \in H_1$, $|\theta\rangle \in H_2$ and all observables act either on H_1 or H_2. In this case the linear combinations $|\eta\rangle = \alpha|\psi\rangle + \beta|\theta\rangle$ have no physical meaning in the following sense:

$$\langle\eta|A|\eta\rangle = \mathrm{trace}(\sigma A)$$

for all A, where $\sigma = |\alpha|^2|\psi\rangle\langle\psi| + |\beta|^2|\theta\rangle\langle\theta|$, so the artificial superposition might as well be replaced by a density matrix. A thorough discussion of superselection

rules can be found in Giulini (2009).

19 Quantum statistical inference

It is an important task now to depart from the assumption of perfect measurements, and address measurements with real data z. I first introduce the concept of operator-valued measure.

Assume a likelihood $p(z^b|\lambda^b = u_j^b)$ for the e-variable λ^b. Define an operator-valued measure M by $M(\{z^b\}) = \sum_j p(z^b|\lambda^b = u_j^b)|b;j\rangle\langle b;j|$. These operators satisfy $M(S) = I$ for the whole sample space and are countably additive. Let the current state be given by the question: 'What is the value of λ^a?', and then the probabilities $\pi^a(u_k^a)$ for the different values $\lambda^a = u_k^a$. Then, defining $\sigma = \sum_k \pi^a(u_k^a)|a;k\rangle\langle a;k|$ we get $P(B) = \text{trace}(\sigma M(B))$ for all sets B in the sample space. This is again proved by a straightforward argument from Born's formula.

In some applications the density matrix σ depends upon an unknown parameter θ. Then the probability measure P above also depends upon θ, and we obtain a statistical model. This is the point of departure of Barndorff-Nielsen et al. (2003), where many notions of ordinary statistical inference theory are generalized.

Related to this is the phenomenon of collapse of the wave packet. Assume first an initial state $|a;k\rangle$, and then an ideal measurement giving the value $\lambda^b = u_j^b$. After the measurement the state then changes to $|b;j\rangle$. This discontinuous change of the state has been considered a great problem in the very common general ontological view on quantum mechanics, problems so great that some physicists adhere to a many-worlds interpretation (see Everett 1973) to cope with it. In our statistical interpretation the collapse represents no problem. A similar 'collapse' occurs in Bayesian statistics once an observation is made.

The situation is similar, but more complicated when a real measurement is made. To cope with this, Barndorff-Nielsen et al. (2003) introduced the notion of an *instrument*. A simple instrument is one where the state is transformed by projecting onto orthogonal subspaces of the Hilbert space, together spanning the whole space. This is also called the Lüders-von Neumann projection postulate, and is similar to the collapse in the ideal measurement above. It is indicated in op. cit. that more general instruments can be formed by combining this with Schrödinger evaluation (see Section 23) and forming compound systems.

There is a large literature on quantum statistical inference. The field started with the monographs of Helstrom (1976) and Holevo (1982), the latter continued in Holevo (2001). There is much more material in Barndorff-Nielsen et al. (2003). Hayashi (2005) is a collection of papers on the asymptotic theory of quantum statistical inference.

20 Entanglement, EPR and the Bell theorem

The total spin components in different directions for a system of two spin 1/2 particles satisfy the assumptions of a maximal symmetric epistemic setting. Assume that we have such a system where $j = 0$, that is, the state is such that the total spin is zero. By ordinary quantum mechanical calculations, this state can be explicitly written as

$$|0\rangle = \frac{1}{\sqrt{2}}(|1,+\rangle \otimes |2,-\rangle - |1,-\rangle \otimes |2,+\rangle), \tag{10}$$

where $|1,+\rangle \otimes |2,-\rangle$ is a state where particle 1 has a spin component $+1/2$ and particle 2 has a spin component $-1/2$ along the z-axis, and *vice versa* for $|1,-\rangle \otimes |2,+\rangle$. This is what is called an entangled state, that is, a state which is not a direct product of the component state vectors. I will follow my own programme, however, and stick to the e-variable description.

Assume further that the two particles separate, the spin component of particle 1 is measured in some direction by an observer Alice, and the spin component of particle 2 is measured by an observer Bob. Before the experiment, the two observers agree both either to measure spin in some fixed direction a or in another fixed direction b, orthogonal to a, both measurements assumed for simplicity to be perfect. As a final assumption, let the positions of the two observers at the time of measurement be spacelike, that is, the distance between them is so large that no signal can reach from one to the other at this time, taking into account that signals cannot go faster that the speed of light by the theory of relativity.

This is Bohm's version of the situation behind the argument against the completeness of quantum mechanics as posed by Einstein et al. (1935) and countered by Bohr (1935 a, b). This discussion is still sometimes taken up today, although most physicists now support Bohr.

I will be very brief on this discussion here. Let λ be 2 times the spin component as measured by Alice, and let η be 2 times the spin component as measured by Bob. Alice has a free choice between measuring in the the directions a and in the direction b. In both cases, her probability is 1/2 for each of $\lambda = \pm 1$. If she measures $\lambda^a = +1$, say, she will predict $\eta^a = -1$ for the corresponding component measured by Bob. According to Einstein et al. (1935) there should then be an element of reality corresponding to this prediction, but if we adapt the strict interpretation of Section 10 here, there is no way in which Alice can predict Bob's actual real measurement at this point of time. Bob on his side has also a free choice of measurement direction a or b, and in both cases he has the probability 1/2 for each of $\eta = \pm 1$. The variables λ and η are conceptual, the first one connected to Alice and the second one connected to Bob. As long as the two are not able to communicate, there is no sense in which we can make statements like $\eta = -\lambda$ meaningful.

The situation changes. however, if Alice and Bob meet at some time after the measurement. If Alice then says 'I chose to make a measurement in the direction a and got the result u' and Bob happens to say 'I also chose to make a measurement in the direction a, and then I got the result v', then these

two statements must be consistent: $v = -u$. This seems to be a necessary requirement for the consistency of the theory. There is a subtle distinction here. The clue is that the choices of measurement direction both for Alice and for Bob are free and independent. The directions are either equal or different. If they should happen to be different, there is no consistency requirement after the measurement, due to the assumed orthogonality of a and b.

Let us then look at the more complicated situation where a and b are not necessarily orthogonal, where Alice tosses a coin and measures in the direction a if head and b if tail, while Bob tosses an independent coin and measures in some direction c if head and in another direction d if tail. Then there is an algebraic inequality

$$\lambda^a \eta^c + \lambda^b \eta^c + \lambda^b \eta^d - \lambda^a \eta^d \leq 2. \tag{11}$$

Since all the conceptual variables take values ± 1, this inequality follows from

$$(\lambda^a + \lambda^b)\eta^c + (\lambda^b - \lambda^a)\eta^d = \pm 2 \leq 2.$$

Now replace the conceptual variables here with actual measurements. Taking then formal expectations from (11), assumes that the products here have meaning as random variables; in the physical literature this is stated as an assumption of realism and locality. This leads formally to

$$E(\widehat{\lambda^a}\widehat{\eta^c}) + E(\widehat{\lambda^b}\widehat{\eta^c}) + E(\widehat{\lambda^b}\widehat{\eta^d}) - E(\widehat{\lambda^a}\widehat{\eta^d}) \leq 2 \tag{12}$$

This is one of Bell's inequalities, called the CHSH inequality.

On the other hand, using quantum-mechanical calculations, that is Born's formula, from the basic state (10), shows that a, b, c and d can be chosen such that Bell's inequality (12) is broken. This is also confirmed by numerous experiments with electrons and photons.

From our point of view the transition from (11) to (12) is not valid. One can not take the expectation term by term in equation (11). The λ's and η's are conceptual variables belonging to different observers. Any valid statistical expectation must take one of these observers as a point of departure. Look at (11) from Alice's point of view, for instance. She starts by tossing a coin. The outcome of this toss leads to some parameter λ being measured in one of the directions a or b. This measurement is an epistemic process, and any prediction based upon this measurement is a new epistemic process. The concept of ancillary statistic from Subsection 6.2 generalizes immediately from inference on a parameter to prediction or inference on any e-variable. In particular, the outcome of the coin toss here is ancillary. By the conditionality principle GPC of Section 7 (the case 1) there) in any epistemic process for Alice, she should condition upon this ancillary. So in any prediction she should condition upon the choice a or b.

By doing predictions from this result, she can use Born's formula. Suppose that she measures λ^a and finds $\lambda^a = +1$, for instance. Then she can predict the value of λ^c and hence $\eta^c = -\lambda^c$. Thus she can (given the outcome a of

the coin toss) compute the expectation of the first term (11). similarly, she can compute the expectation of the last term in (11). But there is no way in which she simultaneously can predict λ^b and η^d. Hence the expectation of the second term (and also, similarly the third term) in (11) is for her meaningless. A similar conclusion is reached if the outcome of the coin toss gives b. And of course a similar conclusion is valid if we take Bob's point of view. Therefore the transition from (11) to (12) is not valid, not by non-locality, but by a simple use of the conditionality principle. This can also in some sense be called lack of realism: In this situation is it not meaningful to take expectation from the point of view of an impartial observer. By necessity one must see the situation from the point of view of one of the observers Alice or Bob.

Entanglement is very important in modern applications of quantum mechanics, not least in quantum information theory, including quantum computation. It is also an important ingredient in the theory of decoherence (Schlosshauer, 2007), which explains why ordinary quantum effects are not usually visible on a larger scale. Decoherence theory shows the importance of the entanglement of each system with its environment. In particular, it leads in effect to the conclusion that all observers share common observations after decoherence between the system and its environment, and this can then be identified with the 'objective' aspects of the world; which is also what the superior actor D of Section 16 would find.

21 Mermin's experiment

Mermin (1985) discusses the following hypothetical experiment to illustrate the peculiar features of quantum mechanics:

Two detectors, one belonging to Alice and one belonging to Bob, are far from each other, and no communication between the two detectors is permitted. Each detector has a switch that can be set in one of three positions 1, 2 or 3, and each detectors responds to an event by either flashing a green light (G) or flashing a red light (R). Midway between the two detectors there is a source emitting particles, causing simultaneous events at the two detectors.

Alice chooses her switch randomly and so does Bob with his switch. A third observer reads off the positions of the switches and the responses R or G after each event. For instance, 32RG means that Alice has position 3, Bob position 2, there is a red flash at Alice's detector and a green flash at Bob's detector. A large number of events of this type is read off. After this, the observer notes the following:

1) If one examines only those runs in which both the switches have the same setting, then one finds that the lights always flash the same colors.

2) If one examines all runs, without any regard to how the switches are set, then one finds that the pattern of flashing is completely random. In particular, half the time the lights flash the same color, and half the time different colors.

Is this really possible? asks Mermin, and answers that by classical thinking it is not. Imagine that the detectors are triggered by particles that have a common origin at the source. Suppose, for example, that what each particle encounters as it enters one detector is a target divided into eight regions, labeled RRR, RRG, RGR, RGG, GRR, GRG, GGR, and GGG. Suppose that each detector is wired so that if a particle lands in the GRG bin, the detector flips into a mode in which the light flashes G if the switch is set to 1, R if it is set to 2, and G if it is set to 3; RGG leads to a mode with R for 1 and G for 2 and 3, and so on. One can imagine variants of this, but all such variants leads to an instruction set of this type. The feature 1) will then result for all possible switch settings of the two detectors if and only if both Alice's detector and Bob's detector receives the same instruction.

Can this then be made consistent with the observation 2)? The answer is no. For the purpose of the present argument one can let the probability of each of RRR, RRG,... be arbitrary. Given that the result is RRG, then the detectors will flash the same color when the switches are set to 11, 22, 33, 12, or 21; they will flash different colors for 13, 31, 23. or 32. Thus with this result, the detectors will flash the same color 5/9 of the time. With exactly the same reasoning, for all the results RRG, GRR, RGR, RGG, GRG, and GGR, the detectors will flash the same color 5/9 of the time, since this argument only depends upon the fact that one color appears twice and the other once. But in the remaining cases RRR and GGG, the detectors always flash the same color. Thus by classical thinking, the two detectors will by necessity flash the same color at least 5/9 of the time. This is inconsistent with 2).

The argument just given, corresponds to Bell's inequality for this experiment. The point is now that according to quantum mechanics, Bell's inequality is violated: One can indeed make a quantum mechanical experiment in which both 1) and 2) holds!

Let the source produce two particles of spin 1/2 in the singlet state, that is, with the total spin equal to 0. Let Alice's detector be as follows: If the switch has position 1, she asks for the spin component in the z-direction; if the switch has position 2 or 3, she asks for the spin component in two different directions in a plane orthogonal to the line towards the source, each separated 120^o from the z-axis. The detector flashes green if the answer is $+1/2$, red if the answer is $-1/2$. Bob's detector is similar, except that position 1 corresponds to a question in the $-z$-direction, and the directions of his positions 2 and 3 are opposite to Alice's directions corresponding to positions 2 and 3. With this arrangement, it is obvious that 1) will hold always. A straightforward calculation using Proposition 5 shows that 2) also holds.

Thus one must by necessity conclude that the classical argument does not hold in the quantum-mechanical setting. What is wrong? According to my view, one must take into account that there are different observers here, and the classical argument must be replaced by an argument from the point of view of one of the observers. First, let us take Alice's point of view. As in the previous section, a valid argument must be conditional on the position chosen by her switch. Also, everything should be conditional on the context. Our task

is to find some context where the quantum-mechanical result can be explained.

We describe the context in terms of a third actor Charles, which must be assumed to act before any event is observed and before Alice and Bob make their choices. The actions of Charles will be described in very concrete terms. He is assumed to have a box containing 4 balls, three yellow and one blue. Before each event he draws a ball randomly from the box. This is the background for producing the results of Alice and Bob. The context is then such that Alice and Bob always get the same result if their switch is the same. If Alice and Bob have different switches, the context give them the same result if the ball chosen by Charles is blue, opposite if the ball is yellow. The whole procedure is repeated for every event.

So let us first look at the experiment from Alice's point of view. To be concrete, assume that she chooses switch 1 and gets as her result a green flash. She does not know Bob's switch position, but she knows the context. Thus she knows that if then Bob chooses switch 1, his result will also be a green flash. If he chooses switch 2 or 3, his probability of green will be 1/4 and probability of red 3/4. If the switch is not recorded, his probability of green will be $1/3 \cdot 1 + 2/3 \cdot 1/4 = 1/2$. Thus the experiment will satisfy 2). She knows from the context that it satisfies 1).

The situation is similar from Bob's point of view.

This situation can also be looked upon by an impartial observer David predicting only one event, but knowing the context. He can make predictions under two circumstances: a) knowing the two switch positions or b) not knowing the switch positions. In the last case he will predict according to 2) that the probability of equal flash is equal to 1/2 by the same argument as used for Alice. In the circumstance a) he will predict equal flash if the switch positions are equal, otherwise a probability 1/4 of equal flash. If he makes predictions for several events, but staying all the time in the same state, the results will be the same.

We need not worry how nature chooses the actor Charles. It is only necessary that one such context produces the result of the quantum experiment. The main thing is the use of the statistical conditionality principle: The result of the whole experiment should be conditional, given the ancillary knowledge of the observer.

In my view, Mermin's hypothetical experiment clarifies the role of the Bell type inequalities, and the reason why such inequalities can be violated in quantum mechanics.

22 The free will theorem

Throughout the times, several authors have proposed various types of hidden variable theories which they claim to be consistent with quantum mechanics. Again and again the scopes of these theories have been limited by so-called no-go-theorems. One of the first and most well-known of these theorems was that of Kochen and Specker (1967): If a theory should be compatible with quantum mechanics, one can not find an arbitrary set of hidden variables that are non-contextual and take definite values at any time.

The newest no-go-theorem is The Free Will Theorem of Conway and Kochen (2006, 2008). They take as a point of departure the EPR-type experiment with spin 1 particles, but presumably this can be generalized. They state two assumptions that are weaker than, but implied by quantum mechanics and one assumption which is implied by relativity theory. Under these assumptions they prove:

THE FREE WILL THEOREM. *If the choice of directions in which to perform spin 1 experiments is not a function of the information accessible to the experimenters, then the responses of the particles are equally not functions of the information accessible to them.*

Thus the particles in a sense have a free will: Their responses are not in any way determined by past history. Past history is here a very wide concept. It can include stochastic variables given in advance, so this kind of simple randomness will not help.

The specific assumptions that Conway and Kochen give for their free will theorem are:

1) SPIN.

Measurements of Alice and Bob are both given in some frame (x, y, z), and the measurements are always 1, 0,1 in some order. This is in particular satisfied by the squares of the spin 1 components along the coordinate axes according to quantum mechanics, which are commuting operators.

2) TWIN.

If the measurements performed by Alice and Bob are along the same axis, they give the same result. This is analogous to what we assumed in the Bell experiment discussed above, only that the signs of Bob's measurements are reversed.

3) FIN.

There is a finite upper bound to the speed with which information can be effectively transmitted. This assumption is weakened in Conway and Kochen (2008).

Admittedly, especially the SPIN-assumption describes a rather special situation, but one can assume that the theorem can be generalized to other situations with entanglement, in the language of the present book: To other situations where Alice and Bob choose their measurement freely, but in different contexts. The result of the Free Will Theorem is then that Nature also chooses its response freely: It is not in any way a function of the past history of the universe.

23 The Schrödinger equation

During a time when no measurement is done on the system, the ket vector is known in quantum mechanics to develop according to the Schrödinger equation:

$$i\hbar \frac{d}{dt}|\psi\rangle_t = H|\psi\rangle_t, \tag{13}$$

where H is a selfadjoint operator called the Hamiltonian (the total energy operator).

I will give two sets of arguments for the Schrödinger equation, one rough and general, and then one specific related to position. The last argument also includes a discussion of the wave function.

23.1 The general argument

Assume that the system at time 0 has a context given by the ket $|\psi\rangle_0$ and at time t by the ket $|\psi\rangle_t$. Let us assume that we ask an epistemic question about the variable θ, and that the ket corresponding to a specific value of this variable is $|\theta\rangle_0$ at time 0 and $|\theta\rangle_t$ at time t. We have the choice between making an ideal measurement at time 0 or at time t. Since there is now disturbance through measurement of the system between these two time points, the probability distribution of the answer must be the same whatever choice is made. Hence according to Born's formula

$$|_0\langle\theta|\psi\rangle_0|^2 = |_t\langle\theta|\psi\rangle_t|^2. \tag{14}$$

Now we refer to a general theorem by Wigner (1959), proved in detail by Bargmann (1964): If an equation like (14) holds, then there must be a unitary or antiunitary transformation from $|\psi\rangle_0$ to $|\psi\rangle_t$. (Antiunitary U means $U^{-1} = -U^\dagger$.) Since by continuity an antiunitary transformation can be excluded here, so we have

$$|\psi\rangle_t = U_t|\psi\rangle_0$$

for some unitary operator U_t. Writing $U_t = \exp(\frac{A_t}{ih})$ for some selfadjoint operator A_t, and assuming that A_t is linear in t: $A_t = Ht$, this is equivalent to (13).

23.2 Position as an inaccessible stochastic process

As in Section 15 consider the motion of a non-relativistic one-dimensional particle, but now make time explicit. Since momentum and hence velocity cannot be determined simultaneously with arbitrary accuracy, it is also impossible to determine positions $\xi(s)$ and $\xi(t)$ simultaneously for two different time points s and t with arbitrary accuracy. Hence the vector $(\xi(s), \xi(t))$ is inaccessible. Fix a time point t. Different observers may focus on different aspects from the past of the time t in order to try to predict $\xi(t)$ as well as possible. These aspects may be formulated by propositional logic in different ways, but for reasons discussed in Appendix 6 I will in this book concentrate on a probabilistic description. Thus observer i may predict $\xi(t)$ by conditioning on some σ-algebra $\mathcal{P}_i$ of information from the past. This may be information from some specific time point s_i with $s_i < t$, but it can also take other forms. We must think of these different observers as hypothetical; only one of them can be realized. Nevertheless one can imagine that all possible information, subject to the choice of observer later, is collected in an inaccessible σ-algebra $\mathcal{P}_t$, the past of $\xi(t)$. The distribution

of $\xi(t)$, given the past $\mathcal{P}_t$, for each t, can then be represented as a stochastic process.

In the simplest case one can then imagine $\{\xi(s); s \geq 0\}$ as an inaccessible Markov process: The future is independent of the past, given the present. Under suitable regularity conditions, a continuous Markov process will be a diffusion process, i.e., a solution of a stochastic differential equation of the type

$$d\xi(t) = b(\xi(t), t)dt + \sigma(\xi(t), t)dw(t). \tag{15}$$

Here $b(\cdot, \cdot)$ and $\sigma(\cdot, \cdot)$ are continuous functions, also assumed differentiable, and $\{w(t); t \geq 0\}$ is a Wiener process. The Wiener process is a stochastic process with continuous paths, independent increments $w(t) - w(s)$, $w(0) = 0$ and $E((w(t) - w(s))^2) = t - s$. Many properties of the Wiener process have been studied, including the fact that its paths are nowhere differentiable. The stochastic differential equation (15) must therefore be defined in a particular way; for an introduction to Itô calculus or Stochastic calculus; see for instance Klebaner (1998). One well known result is Itô's formula: For a two times continuously differentiable function f one has:

$$df(\xi(t), t) = f_t(\xi(t), t)dt + f_x(\xi(t), t)d\xi(t) + \frac{1}{2}f_{xx}(\xi(t), t)\sigma^2(\xi(t), t)dt. \tag{16}$$

There is also the Fokker-Planck equation for the probability density $\rho(x, t)$ of $\xi(t)$:

$$\rho_t(x, t) = -(b(x, t)\rho(x, t))_x + \frac{1}{2}(\sigma^2(x, t)\rho(x, t))_{xx}.$$

So far we have considered observers making predictions of the present value $\xi(t)$, given the past $\mathcal{P}_t$. There is another type of epistemic processes which can be described as follows: Imagine an actor A which considers some future event for the particle, lying in a σ-algebra $\mathcal{F}_j$. He asks himself in which position he should place the particle at time t as well as possible in order to have this event fulfilled. In other words, he can adjust $\xi(t)$ for this purpose. Again one can collect the σ-algebras for the different potential actors in one big inaccessible σ-algebra $\mathcal{F}_t$, the future after t. The conditioning of the present, given the future, defines $\{\xi(t); t \geq 0\}$ as a new inaccessible stochastic process, with now t running backwards in time. In the simplest case this is a Markov process, and can be described by a stochastic differential equation

$$d\xi(t) = b_*(\xi(t), t)dt + \sigma_*(\xi(t), t)dw_*(t), \tag{17}$$

where again $w_*(t)$ is a Wiener process.

Since t is now running backwards in time, Itô's formula now reads:

$$df(\xi(t), t) = f_t(\xi(t), t)dt + f_x(\xi(t), t)d\xi(t) - \frac{1}{2}f_{xx}(\xi(t), t)\sigma_*^2(\xi(t), t)dt. \tag{18}$$

The Fokker-Planck equation is now:

$$\rho_t(x, t) = -(b_*(x, t)\rho(x, t))_x - \frac{1}{2}(\sigma_*^2(x, t)\rho(x, t))_{xx}.$$

23.3 Nelson's stochastic mechanics

Without having much previous knowledge about modern stochastic analysis and without knowing anything about epistemic processes. Nelson (1967) formulated his stochastic mechanics, which serves our purpose perfectly. Nelson considered the multidimensional case, but for simplicity, I will here only discuss a one-dimensional particle. Everything can be generalized.

Nelson discussed what corresponds to the stochastic differential equations (15) and (17) with σ and σ_* constant in space and time. Since heavy particles fluctuate less than light particles, he assumed that these quantities vary inversely with mass m, that is, $\sigma^2 = \sigma_*^2 = \hbar/m$. The constant $\hbar$ has dimension action, and turns out to be equal to Planck's constant divided by 2π. This assumes that $\sigma^2 = \sigma_*^2$, a fact that Nelson actually proved in addition to proving that

$$b_* = b - \sigma^2(\ln\rho)_x.$$

Now define

$$u = \frac{1}{2}(b - b_*), \quad v = \frac{1}{2}(b + b_*).$$

Then

$$u = \frac{1}{2}\sigma^2(\ln\rho)_x,$$

and the two Fokker-Planck equations give the continuity equation

$$\rho_t = -(v\rho)_x.$$

By a simple manipulation from this, one finds that

$$u_t = -\frac{1}{2}\sigma^2 v_{xx} - (vu)_x. \tag{19}$$

Related to (16) with (15) inserted and (18) with (17) inserted, Nelson defined the forward and backward derivatives

$$Df(x(t),t) = f_t(x(t),t) + b(x(t),t)f_x(x(t),t) + \frac{1}{2}\sigma^2 f_{xx}(x(t),t);$$

$$D_*f(x(t),t) = f_t(x(t),t) + b_*(x(t),t)f_x(x(t),t) - \frac{1}{2}\sigma^2 f_{xx}(x(t),t),$$

and argued that the acceleration of the particle can be defined by

$$a(t) = \frac{1}{2}D_*Dx(t) + \frac{1}{2}DD_*x(t).$$

Then a simple manipulation shows that $D\xi(t) = b(\xi(t),t)$, $D_*\xi(t) = b_*(\xi(t),t)$ and that

$$v_t = a + uu_x - vv_x + \frac{1}{2}\sigma^2 u_{xx}. \tag{20}$$

By Newton's law, the force F upon the particle is ma. Assuming that F is derived as the negative gradient of a potential V, we get $a = -m^{-1}V_x$. Inserting

this and at the same time $\sigma^2 = \hbar/m$ into (19) and (20), we have a coupled nonlinear set of differential equations for $u(x,t)$ and $v(x,t)$. This can be solved as an initial value problem assuming $u(x,0) = u_0(x)$ and $v(x,0) = v_0(x)$ for some given functions u_0 and v_0.

From the relationship between b and b_* we already know that

$$R_x = \frac{m}{\hbar}u,$$

where $R(x,t) = \frac{1}{2}\ln\rho(x,t)$. Let S be defined up to an additive constant by

$$S_x = \frac{m}{\hbar}v,$$

and define the complex function $f(x,t)$ by

$$f = e^{R+iS}.$$

Then $|f(x,t)|^2 = \rho(x,t)$. Nelson interpreted f as the wave function of the particle.

A remarkable fact, noted by Nelson, is that the nonlinear set of equations (19) and (20) for u and v transforms into a linear equation

$$f_t = i\frac{\hbar}{2m}f_{xx} - i\frac{1}{\hbar}Vf + i\alpha(t)f. \tag{21}$$

To prove this, we compute the derivatives in (21) and divide by f, finding

$$R_t + iS_t = i\frac{\hbar}{2m}(R_{xx} + iS_{xx} + [R_x + iS_x]^2) - i\frac{1}{\hbar}V + i\alpha(t).$$

Taking x-derivatives here and separating real and imaginary parts, we see that this is equivalent to the pair of equations

$$u_t = -\frac{\hbar}{2m}v_{xx} - (vu)_x,$$

$$v_t = \frac{\hbar}{2m}u_{xx} + \frac{1}{2}(u^2)_x - \frac{1}{2}(v^2)_x - \frac{1}{m}V_x.$$

This is the same as (19) and (20).

Finally, Nelson notes that since the integral of ρ is 1, hence independent of t, if (21) holds at all then $\alpha(t)$ must be real. By choosing, for each t, the arbitrary constant in S appropriately, we can arrange for $\alpha(t)$ to be 0. Thus (21) is equivalent to

$$i\hbar f_t(x,t) = \frac{1}{2m}[(-i\hbar\frac{\partial}{\partial x})^2 + V(x)]f(x,t). \tag{22}$$

This is the Schrödinger equation (13) with the Hamiltonian corresponding to the sum of kinetic and potential energy. Note that as in Section 15 the operator for the momentum of the particle is $-i\hbar\frac{\partial}{\partial x}$.

As already noted, the argument here can be generalized to a multidimensional particle, and also to a system of particles. An open problem at present is to connect it to the derivation of the Hilbert space which was given in Section 11 and the following sections here. My conjecture is that this can only be done if the whole set of arguments is generalized to the relativistic case. A possible starting point for such a generalization may be the paper by Wigner (1939) on the unitary representation of the inhomogeneous Lorentz group and the theory that has been developed from this paper. If such a generalization could be done, it might well supplement the theory of quantum electrodynamics with its renormalization. (An account of quantum electrodynamics together with its history is given in Schweber (1994). The basis for Richard Feynman's derivation of the theory can be found in Brown (2005).)

24 Discussion

This paper falls naturally into two parts, Sections 1-9 on conventional inference, and the last sections on quantum mechanics, although the two parts are closely tied together. Let us first discuss some of the results of the first part.

As indicated in Section 9, statistical inference is often made in steps. At each step, the results of the previous steps then form a part of the context. And it may initiate more steps. A typical case is when a least squares estimation is done in multiple regression, and this is followed by a residual analysis. Such a sequence is not consistent with the ordinary likelihood principle, but it is consistent with our extended basis.

It is often stated that a weakness with the definitions of sufficiency and ancillarity is that they are strongly model dependent. This can remedied by a stepwise analysis, where new models are tested in steps, following a residual analysis from older models.

Taking more steps in the total inference, our basis is even consistent with algorithmic procedures like Breiman's trees (see Breiman, 2001, and references there) and the ordinary partial least squares algorithm (Martens and Næs, 1989).

All this must be taken under one proviso, however: The overall goal of the statistical analysis must be formulated first, and taken as part of the context for all the steps.

I have not gone much into formal logic in this paper. In Section 5 I indicated an equivalence between propositional logic and the ordinary basis for probability models. When I now discuss inference in steps, the propositional logic must be extended to temporal logic, for which there is a large literature; see an introduction in Venema (2001). A further extension would be to proceed to first order and higher order logic, but this is beyond the scope of the present paper.

As a transition between the two parts of this paper, the following remark was made: Statistical literature has much discussion about the way to do inference, but very little on the choice of what to do inference about. These different

questions may be conflicting, even complementary. The symmetrical epistemic setting is a way to formalize a situation where only one out of many possible questions may be addressed.

Here are some problems that must be considered open from the point of view of the present approach, even though some of them are formally solved by conventional quantum mechanics:

- First two technical open question: What conditions should a maximal symmetrical epistemic setting satisfy in order that all unit vectors in H are proportional to $|a;k\rangle$ for some a and k? Or in order that all orthonormal sets of unit vectors in H are of the form $\{|a;k\rangle; k = 1, ..., d\}$ except for phase factors? A necessary condition is that no pairs of states are subject to a superselection rule. On the other hand, the first statement holds for spin 1/2 particles, as shown in Helland (2010). It is also easy to show that the second statement holds for this case.

- Can one find examples where one is sure that the rational epistemic setting is indicated in the macroscopic world? If possible, can such examples be used in a constructive way in quantum information theory and practice?

- As discussed in Section 23 one can also treat continuous systems from the point of view an epistemic process, but this treatment is not closely related to the discussion of Section 11 and the following sections. A reconciliation may induce technical problems, but these problems should be solvable given the vast literature in related mathematics and theoretical physics over the years. However, this development may also induce problems of fundamental art.

- What about a discussion of open systems?

- Can the group-theoretical approach used here in some way throw more light upon elementary particle theory, where group theory is used extensively?

- Can a further development of the discussion here, extended to continuous systems, lead to a reconciliation of quantum theory and relativity theory? It is well known that quantum mechanics can be extended to take into account the special theory of relativity, but that there are conceptual difficulties involved in finding a synthesis between conventional quantum theory and the general theory of relativity. Of course I do not have any solution to these difficulties at present. Already here, however, it is tempting to suggest that gravitational fields and related physical quantities are e-variables, and that they are inaccessible inside black holes.

A completely different attempt to find a unified approach to statistics, quantum theory and relativity theory is given by Frieden (1998, 2004).

To emphasize again part of the motivation behind the present book, I cite Hardy and Spekkens (2010): 'Quantum theory is a peculiar creature. It was born as a theory of atomic physics early in the twentieth century, but over time its scope has broadened, to the point where it now underpins all of modern physics with the exception of gravity. It has been verified to extreme high accuracy and has never been contradicted experimentally. Yet despite of its enormous success, there is still no consensus among physicists about what this theory is saying about the nature of reality.'

What I have tried to do here, is to suggest a new foundation, and thereby

a new interpretation, by bringing together basic ideas from statistics and from physics. In the words of John A. Wheeler: 'Science owes more to the clash of ideas than to the steady accumulation of facts.'

The underlying concept of both these sciences is that of an epistemic process: The process of obtaining knowledge about nature from observations. We begin with an epistemic question: 'What is the value of θ?', where θ is some conceptual variable. Then at the end of the process we have some knowledge of θ, in the simplest case complete knowledge; $\theta = u_k$.

Quantum mechanics then emerges from the observation that in some cases the values of two conceptual variables θ^a and θ^b can not be assessed simultaneously with arbitrary accuracy by any human being: The vector (θ^a, θ^b) is inaccessible.

Despite of this fact probabilities of θ^b, given θ^a, can be found from Born's formula. The arguments for this formula, as given in the present book, rely on a superior observer D. The probabilities obtained are then probabilities from D's point of view, which must be regarded as objective probabilities. In this way the ontology of nature is restored, even though my arguments started with a set of epistemic processes.

What I have offered through these arguments, is a new language, which in my view must be taken as a common language for the foundation of statistics and the foundation of physics. I am a strong believer in the thought that there exists a conceptual basis which is common to all empirical scientific cultures. In fact, I believe that there in some way should exist a common set of principles behind all human cultures, and that the idea of such a basis ideally speaking ought to be a part of the context for all human beings.

Given our own context, we are all subject to our free will. Thus we differ in our opinions and in our actions, related to the fact our context differ: We have different history, different background and different basic convictions.

Finally, let me cite St. Paul from 1. Corinthians 13.9: 'For we know in part and we prophecy in part.' In the narrow sense, this may be related to statistical inference and prediction. But in fact, we are all at any time participating in several epistemic processes, where we try to assess the values of conceptual variables and where we also try to predict the future. And this is always done in part, that is, in an incomplete way. Throughout the time that I have been writing this book, I have also felt that I have been taking part in an epistemic process. And I am quite convinced that there is still more to do in the attempts to complete this process.

Aknowledgements

I am grateful to Philip Goyal for inviting me to the workshop on Reconstructing Quantum Theory at the Perimeter Institute in 2009 on the basis of Helland (2008). Gudmund H. Hermansen has done some of the calculations in connection to Example 17. Also, thanks to Arne B. Sletsjøe for discussions in effect leading to Proposition 2, to Erik Alfsen for giving me the preprint Hammond (2011), to Kingsley Jones for making me aware of Wigner (1959) and

of Bargmann (1964) and to Dag Normann for information about propositional logic.

References

Bargmann, V. (1964). Note on Wigner's Theorem on symmetry operations. *Journal of Mathematical Physics* **5**, 862-868.

Barndorff-Nielsen, O. E., Gill, R. D. and Jupp, P. E. (2003). On quantum statistical inference. *Journal of the Royal Statistical Society* B **65**, 775-816.

Barut, A. S. and Raczka, R. (1985). *Theory of Group Representation and Applications.* Warsaw: Polish Scientific Publishers

Basu, B. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association* **72**, 355-366.

Bell, J.S. (1987). *Speakable and Unspeakable in Quantum Mechanics.* Cambridge: Cambridge University Press

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle.* Hayward, CA: Institute of Mathematical Statistics.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* Chichester: Wiley.

Bickel, P. J. and Doksum, K. A. (2001). *Mathematical Statistics. Basic Ideas and Selected Topics.* 2. ed. New Jersey: Prentice Hall.

Bing-Ren, L. (1992). *Introduction to Operator Algebras.* Singapore: World Scientific.

Birnbaum, A. (1962). On the foundation of statistical inference. *Journal of the American Statistical Association* **57**, 269-326.

Bohr, N. (1935a). Quantum mechanics and physical reality. *Nature* **136**, 65.

Bohr, N. (1935b). Can quantum-mechanical description of physical reality be considered complete? *Physical Review* **48**, 696-702.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* New York: Wiley.

Breiman, L. (2001). Statistical modelling: The two cultures. *Statistical Science* **16**, 199-231.

Brody, T. (1993). *The Philosophy Behind Physics.* Edited by Luis de la Pea and Peter Hodgson. Berlin: Springer-Verlag.

Brown, L. M. [Ed.] (2005). *Feynman's Thesis. A New Approach to Quantum Theory.* New Jersey: World Scientific.

Busch, P. (2003). Quantum states and generalized observables: A simple proof of Gleason's Theorem. *Physical Review Letters* **91** (12), 120403.

Casella, G. and Berger, R. L. (1990). *Statistical Inference.* Pacific Grove, California: Wadsworth and Brooks.

Caves, C. M., Fuchs, C. A. and Schack, R. (2002). Quantum probabilities as Bayesian probabilities. *Physical Review* **A65**, 022305.

Chiribella, G., D'Ariano, G. M. and Perinotti, P. (2010). Informational derivation of quantum theory. arXiv: 1011.6451 [quant-ph]

Cochran, W. G. (1977). *Sampling Techniques.* 3. ed. New York: Wiley.

Congdon, P. (2006). *Bayesian Statistical Modelling. 2. ed.* Chichester: Wiley.

Conway, J. and Kochen, S. (2006). The free will theorem. *Foundations of Physics.* **36**, 1441-1473.

Conway, J. and Kochen, S. (2008). The strong free will theorem. arXiv: 0807.3286

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science* **22**, 1-26.

Cook, R. D., Li, B. and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica* **20**, 927-1010.

Cook, R. D., Helland, I. S. and Su, Z. (2012). Envelopes and partial least squares regression. Under revision.

Cox, D. R. (1958). Some problems connected with statistical inference. *Annals of Statistics* **29**, 357-372.

Cox, D. R. (1971). The choice betwee ancillary statistics. *Journal of the Royal Statistical Society. Series B.* **33**, 251-255.

Cox, D. R. (2006). *Principles of Statistical Inference.* Cambridge: Cambridge University Press.

Cox, D. R. and Donnelly, C. A. (2011). *Principles of Applied Statistics.* Cambridge: Cambridge University Press.

Dawid, A. P. (1975). On the concept of sufficiency and ancillarity in the presence of nuisance parameters. *Journal of the Royal Statistical Society. Series B* **37**, 248-258.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B.* **41**, 1-31.

Eaton, M. L. (1989). *Group Invariance in Applications in Statistics.* Institute of Mathematical Statistics and American Statistical Association, Hayward, California.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association* **70**, 892-898.

Efron, B. (1998). R.A. Fisher in the 21st century. *Statistical Science* **13**, 95-122.

Einstein, A., Podolsky, B. and Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete? *Physical Review* **47**, 777-780.

Everett, H. (1973). The theory of the universal wave function. In: DeWitt, B. S. and Graham, N. *The Many-Worlds Interpretation of Quantum Mechanics.* Princeton, NJ.: Princeton University Press.

Fields, C. (2011). Quantum mechanics from five physical assumptions. arXiv: 1102.0740 [quant-ph]

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A* **222**, 309-368.

Fivel, D. I. (2012). Derivation of the rules of quantum mechanics from information-theoretic axioms. *Foundations of Physics* **42**, 291-318.

Fraser, D. A. S. (1956). Sufficient statistics with nuisance parameters. *Annals of Mathematical Statistics* **27**, 838-842.

Frieden, B. R. (1998). *Physics from Fisher Information. A Unification.* Cambridge: Cambridge University Press.

Frieden, B. R. (2004). *Science from Fisher Information. A Unification.* Cambridge: Cambridge University Press.

Fuchs, C. A. (2002). Quantum mechanics as quantum information (and only a little more). In: Khrennikov, A. (ed.): Quantum Theory: reconsideration of Foundations. Växjö: Växjö University Press.

Fuchs, C. A. (2010). QBism, the Perimeter of Quantum Bayesianism. arXiv:1003.5209 [quant-ph].

Fuchs, C. and Peres, A. (2000). Quantum theory needs no interpretation. *Physics Today, S-0031-9228-0003-230-0*; Discussion *Physics Today, S-0031-9228-0009-220-6.*

Fuchs, C. A. and Schack, R. (2011). A quantum-Bayesian route to quantum-state space. *Foundations of Physics* **41**, 345-356.

Giulini, D. (2009). Superselection rules. arXiv: 0710.1516v2 [quant-ph].

Godambe, V. P. (1980). On sufficiency and ancillarity in the presence of a nuisance parameter. *Biometrika* **67**, 155-162.

Halpern, J. Y. (1995). Reasoning about knowledge: A survey. In: Gabbay, D. M., Hogger, C. J. and Robinson, J. A.: *Handbook of Logic in Artificial Intelligence and Logic Programming.* Vol. 4. Epistemic and Temporal Reasoning. Oxford: Oxford University Press.

Hammond, P. J. (2011). Laboratory games and quantum behaviour. The normal form with a separable state space. Preprint.

Hardy, L. (2001). Quantum theory from five reasonable axioms. arXiv: quant-ph/0101012

Hardy, l. and Spekkens R. (2010). Why physics needs quantum foundatiions. arXiv: 1003.5008 [quant-ph].

Harris, B. (1982). Entropy. In: Kotz, S. and Johnson, N. L.: *Encyclopedia of Statistical Sciences.* Hoboken, NJ.: Wiley.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Springer Series in Statistics.

Hayashi, E. (ed) (2005). *Asymptotic Theory of Quantum Statistical Inference. Selected papers.* Singapore: World Scientific.

Helland, I. S. (1990). Partial least squares regression and statistical models. *Scandinavian Journal of Statistics* **17**, 97-114.

Helland, I. S. (1995). Simple counterexamples against the conditionality principle. *The American Statistician* **49**, 351-356. Discussion **50**, 382-386.

Helland, I. S. (2004). Statistical inference under symmetry. *International Statistical Review* **72**, 409-422.

Helland, I. S. (2006). Extended statistical modeling under symmetry; the link toward quantum mechanics. *Annals of Statistics* **34**, 42-77.

Helland, I. S. (2008). Quantum mechanics from focusing and symmetry. *Foundations of Physics* **38**, 818-842.

Helland, I. S. (2010). *Steps Towards a Unified Basis for Scientific Models and Methods.* Singapore: World Scientific.

Helland, I. S., Sæbø, S. and Tjelmeland, H. (2012). Near optimal prediction from relevant components. To appear in *Scandinavian Journal of Statistics.*

Helstrom, C. W. (1976). *Quantum Detection and Estimation Theory.* New York: Academic Press.

Holevo, A. S. (1982). *Probabilistic and Statistical Aspects of Quantum Theory.* Amsterdam: North-Holland.

Holevo, A. S. (2001). *Statistical Structure of Quantum Theory.* Berlin: Springer-Verlag

Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91**, 1343-1370.

Klebaner, F. C. (1998). *Introduction to Stochastic Calculus with Applications.* London: Imperial College Press.

Knapp, A. W. (1986). *Representation Theory of Semisimple Groups.* Princeton, New Jersey: Princeton University Press.

Knorr Cetina, K. (1999). *Epistemic Cultures. How the Sciences Make Knowledge.* Cambridge, Massachusetts: Harvard University Press.

Kochen, S. and Specker, E. P. (1967). The problem of hidden variables in quantum mechanics. *Journal of Mathematics and Mechanics*, **17**, 59-87.

LeCam, L. (1990). Maximum likelihood: an introduction *International Statistical Review*, **58**, 153-171.

Lehmann, E. L. (1959). *Testing Statistical Hypotheses.*Hoboken, NJ.: Wiley.

Lehmann, E. L. (1999). *Elements of Large-Sample Theory.*New York: Springer.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation.*New York: Springer.

Ma, Z.-Q. (2007). *Group Theory for Physicists.* New Jersey: World Scientific.

Martens, H. and Næs, T. (1989). *Multivariate Calibration.*Hoboken, NJ.: Wiley.

Masanes, L. (2010). Quantum theory from four requirements. arXiv: 1004.1483 [quant-ph]

McCullagh, P. and Han, H. (2011). On Bayes's theorem for improper mixtures. *Annals of Statistics*, **39**, 2007-2020.

Mermin, N. D. (1985). Is the moon there when nobody looks? Reality and the quantum theory. *Physics Today*, **38**, 38-47.

Messiah, A. (1969). *Quantum Mechanics, Volume II.* Amsterdam: North-Holland.

Murphy, G. J. (1990). *C*-algebras and Operator Theory.* Boston: Academic Press.

Nelson, E. (1967). *Dynamical Theories of Brownian Motion.* Princeton: Princeton University Press.

von Neumann, J. (1932). *Mathematische Grundlagen der Quantenmechanik.* Berlin: Springer.

Neyman, J. and Scott, E. L. (1948). Consistent estimators based on partially consistent observations. *Econometrica* **16**, 1-16.

Nisticò, G. and Sestito, A. (2011). Quantum mechanics, can it be consistent with locality? *Foundations of Physics* **41**, 1263-1278.

Næs, T. and Helland, I. S. (1993). Relevant components in regression. *Scandinavian Journal of Statistics* **20**, 239-250.

Patterson, D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545-554.

Reid, N. (1995). The roles of conditioning in inference. *Statistical Science* **10**, 138-199.

Robinson, P. M. (1991). Consistent nonparametric entropy-based testing. *Review of Economic Studies* **58**, 437-453.

Schack, R. (2006). Bayesian probability in quantum mechanics. *Proc. Valencia/ ISBA World Meeting on Bayesian Statistics.*

Schlosshauer, M. (2007). *Decoherence and the Quantum-to-Classical Transition.* Berlin: Springer-Verlag.

Schweber, S. S. (1994). *QED and the Men Who Made It: Dyson, Feynman, Schwinger, and Tomonaga.* Princeton, New Jersey: Princeton University Press.

Schweder, T. and Hjort, N. L. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics* **29**, 309-332.

Searle, S. R. (1971). *Linear Models.* New York: Wiley.

Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics.* London: Chapman and Hall, Inc.

Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication.* University of Illinois Press.

Stigler, S. M. (1976). Discussion of "On rereading R.A. Fisher" by L.J. Savage. *Annals of Statistics* **4**, 498-500.

Taraldsen, G. and Lindqvist, B.H. (2010). Improper priors are not improper. *The American Statistician* **64**, 154-158.

Timpson, C. G. (2008). Quantum Bayesianism: A study. *Studies in History an Philosophy of Modern Physics* **39**, 579-609.

Vedral, V. (2011). Living in a quantum world. *Scientific American* **304** (6), 20-25.

Venema, Y. (2001). Temporal logic. In: Goble, L., ed., *The Blackwell Guide to Philosophical Logic.* Hoboken, NJ.: Blackwell.

Walicki, M. (2012), *Introduction to Mathematical Logic.* New Jersey: World Scientific.

Wasserman, L. (2004). *All of Statistics. A Concise Course in Statistical Inference.* New York: Springer-Verlag.

Wigner, E. (1939). On unitary representations of the inhomogeneous Lorentz group. *Annals of Mathematics* **40**, 149-204.

Wigner, E, P. (1959). *Group Theory and its Application to the Quantum Mechanics of Atomic Spectra.* New York: Academic Press.

Wijsman, R. A. (1990). *Invariant Measures on Groups and Their Use in Statistics.* Lecture Notes - Monograph Series **14**. Hayward, California: Institute of Mathematical Statistics.

Xie, M. and Singh, K. (2011). Confidence distributions, the frequentist distribution of a parameter - a review. Submitted.

Zhu, Y. and Reid, N. (1994). Information, ancillarity, and sufficiency in the presence of nuisance parameters. *Canadian Journal of Statistics* **22**, 111-123.

APPENDIX 1

Independence and correlation

The concept of independence is important in this paper, or more particularly, the concept of conditional independence. This can be taken as the basis for the statistical concepts of sufficiency and ancillarity in the form that we use them here. For a unified treatment of these and related concept based upon conditional independence, see Dawid (1979). A further purpose of this Appendix is to illustrate a main observation behind the paper: Inference from empirical data and the ideas behind such inference are of interest across many scientific cultures, and in discussing this, we should allow basic concepts and ideas to be taken from several cultures, not only from the traditional statistical sphere. The idea from this Appendix is taken from Everett (1973), a book which is the basis for one of the most extreme interpretation of quantum mechanics: The many-worlds interpretation.

In elementary statistical textbooks it is often stated that uncorrelatedness does not imply independence. But what seems very difficult to find in the standard statistical literature, is the fact that there exists a simple measure of correlation which is 0 if and only if the joint probability distribution is independent. Although this goes back to Shannons basic paper from 1948; see Shannon and Weaver (1949) and is well known among some statisticians (see Harris (1982), where also the multivariate result is formulated, or Robinson (1991), who uses essentially this concept in testing in econometrics), the result is simple enough and important enough for a general discussion of independence to be introduced explicitly. The present presentation is modified from Everett (1973). For simplicity we develop the result only for discrete probability distributions, and then present it in general.

DEFINITION A. *i) Let* $\{P_i\}$ *be a discrete probability distribution. Then the entropy of that distribution is defined as*

$$H(\{P_i\}) = -\sum_i P_i \ln(P_i)$$

with $0\ln(0) = 0$.

ii) For a joint probability distribution $\{P_{ij}\}$ *with marginals* $\{P_{i\cdot}\}$ *and* $\{P_{\cdot j}\}$ *the correlation is defined as*

$$C = C(\{P_{ij}\}) = H(\{P_{i\cdot}\}) + H(\{P_{\cdot j}\}) - H(\{P_{ij}\}).$$

THEOREM A. C *is always* ≥ 0. *It is equal to* 0 *if and only if* $P_{ij} = P_{i\cdot} \cdot P_{\cdot j}$ *for all* i *and* j.

PROOF. Let $Q_{ij} = P_{ij}/P_{i\cdot}P_{\cdot j}$ if $P_{i\cdot}P_{\cdot j} > 0$, otherwise $Q_{ij} = 1$. (Note that $P_{i\cdot}P_{\cdot j} = 0$ implies $P_{ij} = 0$.) Then $P_{ij} = Q_{ij}P_{i\cdot}P_{\cdot j}$ always, and

$$C = \sum_{ij} P_{ij} \ln(\frac{P_{ij}}{P_{i\cdot}P_{\cdot j}}) = \sum_{ij} P_{i\cdot}P_{\cdot j}Q_{ij}\ln(Q_{ij}).$$

For $x \geq 0$ it is always true that $x\ln(x) > x - 1$; except that one has equality for $x = 1$. Hence

$$C > \sum_{ij} P_{i\cdot}P_{\cdot j}(Q_{ij} - 1) = \sum_{ij} P_{ij} - \sum_{ij} P_{i\cdot}P_{\cdot j} = 0$$

except when all the $Q_{ij} = 1$, that is, the case of independence.

DEFINITION B. *i) Consider any random variable u in the sense defined in Section 2. Let $P(\cdot|\tau)$ be its conditional probability distribution, given the e-variable τ, and let $p(u|\tau)$ be its density with repect to a fixed positive measure μ, to be understood. The conditional entropy of that variable is then*

$$H(u|\tau) = -\int_u p(u|\tau)\ln(p(u|\tau))d\mu.$$

ii) Let u and v be two random variables, and assume that the joint entropy $H(u, v|\tau)$ is finite. Then the conditional correlation between u and v is defined as

$$C(u, v|\tau) = H(u|\tau) + H(v|\tau) - H(u, v|\tau).$$

The entropy is always well-defined as a number in $[0, \infty]$. If τ is a σ-algebra, and a measure is defined on this σ-algebra, then H is only uniquely defined almost surely with respect to this measure. This will be understood in the following.

THEOREM B. *Assume that the joint entropy is finite. Then $C(u, v|\tau) \geq 0$, and u and v are conditionally independent, given τ, if and only if $C(u, v|\tau) = 0$.*

Although the concept of entropy is not used explicitly in this paper, the result above deserve to be mentioned, and also the simple observation that entropy is related to the statistical concept of Kullback-Leibler distance. I also want to recall here the well-known fact that dependence between random variables in general implies no causation. For a simple treatment of this aspect; see Wasserman (2004).

APPENDIX 2

Proof that the generalized likelihood principle follows from the GWCP and the GWSP (Birnbaum's theorem); the discrete case

Let E_1 and E_2 be the two experiments in the generalized likelihood principle, and let E^* be the mixed experiment from the GWCP. On the sample space of E^* define the statistic

$$t(j, z_j) = \begin{cases} (1, z_1^*) & \text{if } j = 1 \text{ and } z_1 = z_1^* \text{ or if } j = 2 \text{ and } z_2 = z_2^* \\ (j, z_j) & \text{otherwise} \end{cases} .$$

I will use the factorization theorem to prove that $t(j, z_j)$ is a sufficient statistic in the mixed experiment E^*. Define

$$h(j, z_j|\tau) = \begin{cases} c & \text{if } (j, z_j) = (2, z_2^*) \\ 1 & \text{otherwise} \end{cases} ,$$

where c is the constant of proportionality between the two likelihoods. Define for both values of j and for all z_j:

$$g(t|\theta, \tau) = g((j, z_j)|\theta, \tau) = f^*((j, z_j)|\theta, \tau),$$

where f^* is the point probability in E^*.

Now for all sample points except $(2, z_2^*)$ (but including $(1, z_1^*)$), we have $t(j, z_j) = (j, z_j)$, so

$$g(t(j, z_j)|\theta, \tau)h(j, z_j|\tau) = g((j, z_j)|\theta, \tau) \cdot 1 = f^*((j, z_j)|\theta, \tau).$$

Such a factorization also holds for $(2, z_2^*)$. Namely, by using the definifions of t, h, g, f^* together with the GWCP, we have

$$g(t(2, z_2^*)|\theta, \tau)h(2, z_2^*|\tau) = g((1, z_1^*)|\theta, \tau)c = f^*((1, z_1^*)|\theta, \tau)c = c\frac{1}{2}f_1(z_1^*|\theta, \tau)$$

$$= c\frac{1}{2}L_1(\theta|z_1^*, \tau) = \frac{1}{2}L_2(\theta|z_2^*, \tau) = \frac{1}{2}f_2(z_2^*|\theta, \tau) = f^*((2, z_2^*)|\theta, \tau).$$

Here L_1 and L_2 are the likelihoods of the two experiments E_1 and E_2, and I have used the premise of the generalized likelihood principle.

Thus by the factorization theorem, $t(j, z_j)$ is a sufficient statistic for θ, and by the WGSP we have that the evidence about θ in E^* given by $(1, z_1^*)$ and $(2, z_2^*)$ are the same. By the WGCP, $(1, z_1^*)$ gives the same evidence as z_1^* in E_1 and $(2, z_2^*)$ gives the same evidence as z_2^* in E_2. Hence these latter evidences must also be the same, and the generalized likelihood principle follows.

APPENDIX 3

Some group theory, operator theory and group representation theory

A group G is defined in mathematics as a set of elements g with a composition g_1g_2 satisfying the axioms: (i) There is a unit e such that $eg = ge = g$ for all g; (ii) For each g there is an inverse g^{-1} such that $g^{-1}g = gg^{-1} = e$; (iii) The composition is associative: $(g_1g_2)g_3 = g_1(g_2g_3)$ for all g_1, g_2, g_3.

The group is Abelian (commutative) if $g_1g_2 = g_2g_1$ for all g_1, g_2.

Important examples of groups are the additive group on the real numbers and the multiplicative group on the positive real numbers.

Most of the groups used in this paper are group actions, that is, transformation groups on some set Φ, even though some must be seen as abstract groups. A transformation g of Φ is any function on ϕ which is one-to-one and onto. These functions can be composed by $(g_1g_2)(\phi) = g_1(g_2(\phi))$, and they have inverses g^{-1}. The existence of a unit and the associative law are automatic. Thus by definition they form a group. For any set Φ the group of all transformations on Φ exists, and is called the automorphism group of Φ. Thus many groups G of this paper may be considered as subgroups of some automorphism group.

An orbit of a transformation group G is a subset of Φ, the set of all ϕ that are transformed from a single element ϕ_0, that is $\{\phi : \ \phi = g\phi_0$ for some $g \in G\}$. The restriction of G to an orbit or to a set of orbits will itself be a group transformation, which again without possible confusion can be called G. Restrictions to orbits of groups on the parameter space were used in connection with model reduction in Section 3 and later. This constraint on model reduction is important if the same transformation group shall be kept during the reduction. Such reductions were important in connection to the symmetrical epistemic setting used in introducing the quantum mechanical perspective.

A group where the only orbit is the full group, is said to be transitive. For a transitive group, each element of Φ can be transformed to each other element by some group action.

The stabilizer of an element $\phi_0 \in \Phi$ is the subgroup H of G such that $h(\phi_0) = \phi_0$ for $h \in H$. If $\phi_1 = g\phi_0$, then $H(\phi_1) = gH(\phi_0)g^{-1}$. For some groups the stabilizer is trivial.

Let in general both the set Φ and the group G be given some topology, both spaces assumed to be locally compact. Then one can under quite general conditions (see Helland (2010) or any mathematical text on this) define in a unique way (except for a multiplicative constant) two positive measures, a left Haar measure μ_G satisfying $\mu_G(gD) = \mu_G(D)$ and a right Haar measure ν_G satisfying $\nu_G(Dg) = \nu_G(D)$ for all $g \in G$ and all Borel sets $D \subseteq G$.

Then turn to invariant measures on the set Φ itself. In mathematical texts, Φ is often itself treated as a group, and then the concepts of Haar measures carry over. But this is not satisfactory for all statistical applications. In Helland

(2010, Subsection 3.3 and Appendix A.2.2) a summary of a way to fix this is given. In that book, group actions were written to the right: $\phi \to \phi g$ so that $\phi(g_1 g_2) = (\phi g_1) g_2$. This is uncommon, but has a certain logical advantage: In the product $g_1 g_2$, we get that g_1 comes first and then g_2.

A left-invariant measure on Φ is then any measure μ satisfying $\mu(g(B)) = \mu(B)$ for any $g \in G$ and for any Borel set $B \subseteq \Phi$, while a right-invariant measure is any measure ν satisfying $\nu(Bg) = \nu(B)$ for all g, B. In Helland (2010, Theorem A1) it was proved that a right invariant measure always exists on a given orbit of G if the stabilizer of one element, hence of all elements, of this orbit is compact. This is the case under weak technical assumptions (proper group actions; see Wijsman, 1990) if G is locally compact. In Helland (2010, Subection 3.3) a list of arguments were given why the right invariant measure should be used as an objective prior in statistics if such a prior is required.

The invariant measure is unique up to a multiplicative scalar if the group action is transitive, otherwise invariant measures can be introduced independently on each orbit. For compact groups and in many other cases the left-invariant measure and the right-invariant measure can be taken as identical. When Φ is compact, the invariant measure can be taken as normalized: $\nu(\Phi) = 1$.

Two groups G and R are homomorphic if there exists a function T from G to R such that $T(g_1 g_2) = T(g_1)T(g_2)$ for all g_1, g_2 and such that $T(e) = e'$, the unit in R. Then also $T(g^{-1}) = T(g)^{-1}$. They are isomorphic if T is one-to-one. Then they may be considered as the same abstract group.

Next let us introduce some basic algebra. A vector space is a group under addition where also multiplication by scalars is defined. In this paper we mainly consider finite-dimensional complex vector spaces, meaning that the scalars are complex numbers and that there exists a set of basis vectors $\boldsymbol{e}_i; i = 1, ..., k$ that are linearly independent: $\sum_i c_i \boldsymbol{e}_i = 0$ implies $c_1 = ... = c_k = 0$. A linear operator A on a vector space is a function from the vector space into itself satisfying

$$A(c_1 \boldsymbol{a}_1 + c_2 \boldsymbol{a}_2) = c_1 A\boldsymbol{a}_1 + c_2 A\boldsymbol{a}_2.$$

By relating it to the basis vectors, a linear operator can always be represented by a square matrix:

$$A\boldsymbol{e}_j = \sum_i \boldsymbol{e}_i D(A)_{ij}.$$

Then if $\boldsymbol{a} = \sum_j \boldsymbol{e}_j a_j$ and $\boldsymbol{b} = A\boldsymbol{a} = \sum_i \boldsymbol{e}_i b_i$, we have $b_i = \sum_j D(A)_{ij} a_j$. Thus if $\underline{a}$ is the coloumn vector of components a_j and similarly for $\underline{b}$, we get $\underline{b} = D(A)\underline{a}$; in complete anology to $\boldsymbol{b} = A\boldsymbol{a}$.

If $\boldsymbol{a}$ is represented by the coloumn vector $\underline{a}$, we define $\boldsymbol{a}^\dagger$ as represented by the row vector $(a_1^*, ..., a_k^*)$, where $*$ denotes complex conjugate. The scalar product $\boldsymbol{a}^\dagger \boldsymbol{b}$ is defined as $\sum_i a_i^* b_i$. This scalar product is linear in the second vector and antilinear in the first, in agreement with the tradition in physics. Mathematicians tend to use a scalar product which is linear in the first vector. This is in effect just a cultural difference, but to outsiders it is annoying.

Two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are orthogonal if $\boldsymbol{a}^\dagger \boldsymbol{b} = 0$. With this interpretation, the basis vectors $\boldsymbol{e}_i$ are automatically pairwise orthogonal and have norm $\|\boldsymbol{e}_i\| \equiv$

$\sqrt{\boldsymbol{e}_i^\dagger \boldsymbol{e}_i} = 1$. In general one can always find many sets of n orthogonal basis vectors in an n-dimensional vector space.

A vector space with the structure above is called an inner product space. This notion can be generalized to infinite-dimensional spaces, having an infinite set of basis vectors. The norm $\|\boldsymbol{a}\| = \sqrt{\boldsymbol{a}^\dagger \boldsymbol{a}}$ induces a metric, hence a topology on this space by $d(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{a} - \boldsymbol{b}\|$. The space is complete in this metric if $\|\boldsymbol{a}_n - \boldsymbol{a}_m\| \to 0$ $(n, m \to \infty)$ implies that there exists an $\boldsymbol{a}$ such that $\|\boldsymbol{a}_n - \boldsymbol{a}\| \to 0$. A complete inner product space is called a Hilbert space. A closed subspace of a Hilbert space is again a Hilbert space. A finite-dimensional inner product space is always complete, hence a Hilbert space.

The identity operator I is defined by $I\boldsymbol{a} = \boldsymbol{a}$, and the multiplication of operators by $(AB)(\boldsymbol{a}) = A(B\boldsymbol{a})$. Then (in the finite-dimensional case) $D(I)$ is diagonal with 1's on the diagonal, and $D(AB) = D(A)D(B)$, ordinary matrix multiplication. An operator A is invertible if there exists an A^{-1} such that $A^{-1}A = AA^{-1} = I$. A finite-dimensional operator A is invertible if and only if $\det(D(A)) \neq 0$; then $D(A^{-1}) = D(A)^{-1}$.

The conjugate of an operator A, $A^\dagger$, is defined by $\boldsymbol{a}^\dagger(A\boldsymbol{b}) = (A^\dagger \boldsymbol{a})^\dagger \boldsymbol{b}$. Slightly different, but equivalent notations for scalar products and conjugates, using kets and bras, were used in the main text. An operator A is called Hermitian if $A^\dagger = A$. An operator V is called unitary if $V^{-1} = V^\dagger$.

An eigenvector $\boldsymbol{v}$ and an eigenvalue λ are solutions of $A\boldsymbol{v} = \lambda \boldsymbol{v}$. An operator A is Hermitian if and only if all its eigenvalues are real-valued. Eigenvectors corresponding to different eigenvalues are then automatically orthogonal. In the k-dimensional Hermitian case there are always sets of k pairwise orthogonal eigenvectors.

A group representation of G is a continuous homomorphism from G to the group of invertible linear operators V on some vector space H:

$$V(g_1 g_2) = V(g_1)V(g_2).$$

It is also required that $V(e) = I$, the identity. This assures that the inverse exists: $V(g)^{-1} = V(g^{-1})$. The representation is unitary if the operators are unitary $(V(g)^\dagger V(g) = I)$. If the vector space is finite-dimensional, we have a representation $D(V)$ on the square, invertible matrices. For any representation V and any fixed invertible operator K on the vector space, we can define a new representation by $W(g) = KV(g)K^{-1}$. One can prove that two equivalent unitary representations are unitarily equivalent, so K can be chosen as a unitary operator.

A subspace H_1 of H is called invariant with respect to the representation V if $u \in H_1$ implies $V(g)u \in H_1$ for all $g \in G$. The null-space $\{0\}$ and the whole space H are trivially invariant; other invariant subspaces are called proper. A group representation V of a group G in H is called irreducible if it has no proper invariant subspace. A representation is said to be fully reducible if it can be expressed as a direct sum of irreducible subrepresentations. A finite-dimensional unitary representation of any group is fully reducible. In terms of a matrix representation, this means that we can always find a $W(g) = KV(g)K^{-1}$

such that $D(W)$ is of minimal block diagonal form. Each one of these blocks will represent an irreducible representation. They are all one-dimensional if and only if G is Abelian. The blocks may be seen as operators on subspaces of the original vector space, the irreducible subspaces. These are important in studying the structure of the group.

A useful result is Schur's Lemma (; see for instance Barut and Raczka, 1985):

Let V_1 and V_2 be two irreducible representations of a group G; V_1 on the space H_1 and V_2 on the space H_2. Suppose that there is a transformation T from H_1 to H_2 such that

$$V_2(g)T(v) = T(V_1(g)v)$$

for all $g \in G$ and $v \in H_1$.

Then either T is zero or it is an isomorphism. Furthermore, if $H_1 = H_2$, then $T = \lambda I$ for some complex number λ.

Let ν be the right and left invariant measure of the space Φ induced by the group G, assuming the two to be equal, and consider the Hilbert space $H = L^2(\Phi, \nu)$. Then the right regular representation of G on H is defined by $U^R(g)f(\phi) = f(\phi g)$ and the left regular representation by $U^L(g)f(\phi) = f(g^{-1}\phi)$. These representations always exist, and they can be shown to be unitary.

If V is an arbitrary representation of a compact group G in H, then there exists in H a new scalar product defining a norm equivalent to the initial one, relative to which V is a unitary representation of G.

For references to some of the vast literature on group representation theory, see Helland (2010, Appendix A.2.4).

APPENDIX 4

Proofs related to quantum states

Proof of Theorem 1 of Section 12.

(i) For each a and for $g^a \in G^a$ define $V(g^a) = U(g_{0a})U(g^a)U(g_{a0})$. Then $V(g^a)$ is an operator on $H = L(\Lambda^0)$, since it is equal to $U(g_{0a}g^a g_{a0})$, and $g_{0a}g^a g_{a0} \in G^0$ by Definition 9a). For a product $g^a g^b g^c$ with $g^a \in G^a$, $g^b \in G^b$ and $g^c \in G^c$ we define $V(g^a g^b g^c) = V(g^a)V(g^b)V(g^c)$, and similarly for all elements of G that can be written as a finite product of elements from the subgroups.

Let now g and h be any two elements in G such that g can be written as a product of elements from G^a, G^b and G^c, and similarly h (the proof is similar for other cases.) It follows that $V(gh) = V(g)V(h)$ on these elements, since the last factor of g and the first factor of h either must belong to the same subgroup or to different subgroups; in both cases the product can be reduced by the definition of the previous paragraph. In this way we see that V is a representation on the set of finite products, and since these generate G by Assumption 2c), and since U, hence by definition V, is continuous, it is a representation of G.

Since different representations of g as a product may give different solutions, we have to include the possibility that V may be multivalued.

(ii) Directly from the proof of (i).

Proof of Theorem 2 of Section 12.

(i) Assume as in Theorem 1 that we have a multivalued representation V of G. Define a larger group G' as follows: If $g^a g^b g^c = g^d g^e g^f$, say, with $g^k \in G^k$ for all k, we define $g'_1 = g^a g^b g^c$ and $g'_2 = g^d g^e g^f$. A similar definition of new group elements is done if we have equality of a limit of such products. Let G' be the collection of all such new elements that can be written as a formal product of elements $g^k \in G^k$ or as limits of such symbols. The product is defined in the natural way, and the inverse by for example $(g^a g^b g^c)^{-1} = (g^c)^{-1}(g^b)^{-1}(g^a)^{-1}$. By Assumption 2c), the group G' generated by this construction must be at least as large as G. It is clear from the proof of Theorem 1 that V also is a representation of the larger group G' on H, now a one-valued representation.

(ii) Again, if $g^a g^b g^c = g^d g^e g^f = g$, say, with $g^k \in G^k$ for all k, we define $g'_1 = g^a g^b g^c$ and $g'_2 = g^d g^e g^f$. There is a natural map $g'_1 \to g$ and $g'_2 \to g$, and the situation is similar for other products and limits of products. It is easily shown that this mapping is a homomorphism.

Proof of Theorem 3 of Section 12.

(i) Consider the case where $g' = g^a g^b g^c$ with $g^k \in G^k$. Then by the proof of Theorem 1:

$$V(g') = U_a U(g^a) U_a^\dagger U_b U(g^b) U_b^\dagger U_c U(g^c) U_c^\dagger = U(g_{0a} g^a g_{a0} g_{0b} g^b g_{b0} g_{0c} g^c g_{c0})$$

$$= U(g^0),$$

where $g^0 \in G^0$. The group element g^0 is unique since the decomposition $g' = g^a g^b g^c$ is unique for $g' \in G'$. The proof is similar for other decompositions and limits of these. By the construction, the mapping $g' \to g^0$ is a homomorphism.

(ii) Assume that $g^0 = e$ and $g' \neq e'$. Since $U(g^0)\tilde{f}(\lambda^0(\phi)) = \tilde{f}(\lambda^0((g^0)^{-1}(\phi)))$, it follows from $g^0 = e$ that $U(g^0) = I$ on H. But then from (i), $V(g') = I$, and since V is a univariate representation, it follows that $g' = e'$, contrary to the assumption.

Proof of Proposition 4 of Section 12.

We have $|a;k\rangle = V(g'_a)|0;k\rangle = V(g'_a)I(\lambda^0(\phi) = u_k)$. Since $\tilde{G}^0$ is transitive on the range of λ^0, there is a $\tilde{g}^{0k} \in \tilde{G}^0$ such that $\tilde{g}^{0k} u_k = u_0$. Then $|0;k\rangle = I(\lambda^a((g^{0k})^{-1}\phi) = u_0) = U(g^{0k})I(\lambda^a(\phi) = u_0) = V(g^{0k})|0;0\rangle$ since $V(g^{0k}) = U_0 U(g^{0k}) U_0^\dagger = U(g^{0k})$. So the conclusion holds with $g'(a,k) = g'_a g^{0k}$.

Proof of Theorem 4 of Section 12.

a) I prove the first statement; the second follows from the proof of the first statement. Without loss of generality consider a system where each e-variable λ takes only two values, say 0 and 1. Otherwise we can reduce to a degerate system with just these two values: The statement $|a;i\rangle = |b;j\rangle$ involves, in addition to λ^a and λ^b, only the two values u_i and u_j. By considering a function of the maximally accessible e-variable (cp. Section 13), we can take one specific value equal to 1, and the others collected in 0. By doing this, we also arrange that both u_i and u_j are 1, so we are comparing the state given by $\lambda^a = 1$ with the state given by $\lambda^b = 1$.

By the definition, $|a;1\rangle = |b;1\rangle$ can be written

$$V(g'_a)U_a I(\lambda^a(\phi) = 1) = V(g'_b)U_b I(\lambda^b(\phi) = 1)$$

for group elements g'_a and g'_b in G'.

Use Theorem 3(i) and find g^0_a and g^0_b in G^0 such that $V(g'_a) = U(g^0_a)$ and $V(g'_b) = U(g^0_b)$. Therefore

$$U(g^0_a)U(g_{0a})I(\lambda^a(\phi) = 1) = U(g^0_b)U(g_{0b})I(\lambda^b(\phi) = 1);$$

$$I(\lambda^a(\phi) = 1) = U(g^0)I(\lambda^b(\phi) = 1) = I(\lambda^b((g^0)^{-1}\phi) = 1) = I((\tilde{g}^0)^{-1}\lambda^b(\phi) = 1),$$

for $g^0 = (g_{0a})^{-1}(g^0_a)^{-1} g^0_b g_{0b} \in G^0$.

Both λ^a and λ^b take only the values 0 and 1. Since the set where $\lambda^b(\phi) = 1$ can be transformed into the set where $\lambda^a(\phi) = 1$, we must have $\lambda^a = F(\lambda^b)$ for some transformation F.

b) follows trivially from a).

***

It is not necessary to assume that V is an irreducible representation of the group G on the Hilbert space H. In general the Hilbert space can be decomposed as $H = H_1 \oplus H_2 \oplus ...$, where G has an irreducible representation on each of the spaces H_i.

Not all vectors in H are necessarily possible state vectors. If A is an operator corresponding to an absolutely conserved quantity like the charge or the total spin of a particle, then linear combinations of eigenvectors of A with different eigenvalues are not possible state vectors (superselection rules).

***

Proof of Theorem 6 of Section 15.

Let $\epsilon > 0$ be given. Find first $a > 0$ so large that $\int_{-\infty}^{-a} |f(\xi)|^2 d\xi$, $\int_a^{\infty} |f(\xi)|^2 d\xi$, $\int_{-\infty}^{-a} |\xi f(\xi)|^2 d\xi$, $\int_a^{\infty} |\xi f(\xi)|^2 d\xi$ all are less than $\epsilon/4$. Assume that n is so large that $\xi_{n1} < -a$ and $\xi_{nk_n} > a$. Since f is uniformly continuous on $[-a, a]$ it follows that $\int_{-a}^{a} |f_n(\xi) - f(\xi)|^2 d\xi \to 0$, so $\|f_n - f\| \to 0$. Since $1 - \sum_j I_{nj}$ is less than the indicator of $(-\infty, -a]$ plus the indicator of $[a, \infty)$, we have $\int |\xi f(\xi) - \xi f(\xi) \sum_j I_{nj}(\xi)|^2 d\xi < \epsilon/2$. Now

$$\xi f(\xi) \sum_j I_{nj}(\xi) - A_n f_n(\xi) = \sum_j (\xi f(\xi) - \xi_{nj} f_n(\xi_{nj})) I_{nj}(\xi).$$

Hence using the uniform continuity of $k(\xi) = \xi f(\xi)$ on $[-a, a]$, we get $\int |\xi f(\xi) - A_n f_n(\xi)|^2 \to 0$.

APPENDIX 5

Proof of Busch's Theorem for the finite-dimensional case

The main point of the proof is to show that any generalized probability measure on effects extends to a unique positive linear functional on the vector space of all bounded linear Hermitian operators. This is done in steps.

1) It is trivial that $\mu(E) = n\mu(\frac{1}{n}E)$ for all positive integers. It follows that $\mu(pE) = p\mu(E)$ for all rational numbers in $[0,1]$. By approximating from below and from above by rational numbers, this implies that $\mu(\alpha E) = \alpha\mu(E)$ for all real numbers α in $[0,1]$.

2) Let A be any positive bounded operator in H. Then there is a positive number α such that $\langle u|Au\rangle \leq \alpha$ for all unit vectors u. Then E defined by $E = (1/\alpha)A$ is an effect. Thus we can always write $A = \alpha E$ for an effect E. Assume now that there are two effects E_1 and E_2 such that $A = \alpha_1 E_1 = \alpha_2 E_2$. Assume without loss of generality that $\alpha_2 > \alpha_1 > 0$. Then $\mu(E_2) = \frac{\alpha_1}{\alpha_2}\mu(E_1)$, so $\alpha_1\mu(E_1) = \alpha_2\mu(E_2)$. Therefore we can uniquely define $\mu(A) = \alpha_1\mu(E_1)$.

3) Let A and B be positive bounded operators. Take $\gamma > 1$ such that $\frac{1}{\gamma}(A+B)$ is an effect. Then we can write $\mu(A+B)$ as $\gamma\mu(\frac{1}{\gamma}(A+B)) = \gamma\mu(\frac{1}{\gamma}A) + \gamma\mu(\frac{1}{\gamma}B) = \mu(A) + \mu(B)$.

4) Let C be an arbitrary bounded Hermitian operator. Assume that we have two different decompositions $C = A - B = A' - B'$ into a difference of positive operators. Then $A + B' = A' + B$ implies $\mu(A) + \mu(B') = \mu(A') + \mu(B)$. Hence $\mu(A) - \mu(B) = \mu(A') - \mu(B')$, so we can uniquely define $\mu(C)$ as $\mu(A) - \mu(B)$. It follows then easily from 3) that $\mu(C+D) = \mu(C) + \mu(D)$ for bounded Hermitian operators.

5) This is extended directly to $\mu(C_1 + ... + C_r) = \mu(C_1 + ... + C_{r-1}) + \mu(C_r) = \mu(C_1) + ... + \mu(C_r)$ for finite sums.

Let $\{|k\rangle; k = 1, ..., n\}$ be a basis for H. Then for any Hermitian operator C we can write $C = \sum_{i,j} c_{ij}|i\rangle\langle j|$, where c_{ij} are complex numbers satisfying $c_{ij}* = c_{ji}$. Define the operator σ by $\sigma_{ij} = \mu(|i\rangle\langle j|)$. Then σ is a positive operator since $\langle v|\sigma v\rangle = \mu(|v\rangle\langle v|)$ for any vector $|v\rangle$. Also

$$\text{trace}(\sigma) = \sum_i \sigma_{ii} = \sum_i \mu(|i\rangle\langle i|) = \mu(\sum_i |i\rangle\langle i|) = \mu(I) = 1,$$

so σ is a density operator.

We have $\mu(C) = \sum_{i,j} \sigma_{ij} c_{ij} = \text{trace}(\sigma C)$, and this holds in particular when C is an effect.

APPENDIX 6

Propositional logic, probabilities and knowledge

Mathematical logic can be studies at many different levels. In this book I will concentrate on propositional logic, and I will look at propositions as they are formulated in ordinary, everyday language as primitive entities. For a more formal approach to propositional logic including axioms and a separation between syntax and semantics, see for instance Walicki (2012).

Propositions A and B can be connected: $A \vee B$ means that A or B is true, while $A \wedge B$ means that both A and B are true, similarly for the connection between more propositions. Also, $\neg A$ means that A is not true. We let $\perp$ denote an impossible proposition, while $\top$ denotes a proposition which is aways true. In ordinary texts in mathematical logic one usually works with a finite number of propositions A_i. I will allow for an infinite, even uncountable number of propositions, so that propositions of the form 'The rain tomorrow will amount to less than or equal to x mm' will be permitted for different x.

There is a close connection between propositional logic and set theory. The translation is straightforward: $\vee$ translates into $\cup$, while $\wedge$ translates into $\cap$; $\neg A$ corresponds to A^c, while $\perp, \top$ correspond to $\emptyset, \Omega$, assuming that all the sets are subsets of Ω.

One can also define probabilities of propositions; in fact this is often done in elementary probability texts. With the above translations, there is a close connection to Kolmogorov's axioms; see Subsection 2.1. For instance $P(A_1 \vee A_2 \vee ...) = P(A_1) + P(A_2) + ...$ if the A_i's satisfy $A_i \wedge A_j = \perp$ for each pair. Also $P(\neg A) = 1 - P(A)$. The rule $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$ is always true. It can be proved rigorously, but it can also be motivated by a Venn diagram from the analogue with set theory.

Conditional probabilities can be defined by $P(A|B) = P(A \wedge B)/P(B)$ when $P(B) > 0$. In this book I need the more general notion of conditional probability given a σ-algebra of propositions $\mathcal{B}$, and then it seems like we may need to assume a little more structure. Assume thus that there exists a countable collection of atomic propositions $\{C_i\}$ such that all other propositions A can be formed by combining the C_i's by $\vee$'s, such that $C_i \wedge C_j = \perp$ for pairs and $\top = \bigvee_i C_i$. This assumption simplifies the discussion. It is satisfied in the case of a finite number of propositions closed under $\wedge$ and $\neg$. In general we can think of the C_i's as formed by combining all propositions of interest through $\wedge$'s. The whole σ-algebra $\mathcal{F}$ is generated by the C_i's.

Let now the sub-σ-algebra $\mathcal{B}$ be generated by $\{B_j\}$, partly a subset of $\{C_i\}$ and partly formed by taking $\vee$ of some C_i, such that $B_j \wedge B_k = \perp$ for pairs and such that $\top = \bigvee_j B_j$. Then we can define

$$P(A|\mathcal{B}) = \sum_j P(A|B_j)\mathbf{1}(B_j),$$

where $\mathbf{1}(B_j) = 1$ if B_j is true, 0 if it is not true. From this, $P(A|\mathcal{B})$ is uniquely

defined except on a set with probability 0. The analogue of the Radon-Nikodym definition (1) is then

$$\int_B P(A|\mathcal{B})dP = P(A \wedge B)$$

for all $B \in \mathcal{B}$. One of the open questions of this book is whether this formula can be generalized in the context of propositional logic, and then can be taken as a general definition of $P(A|\mathcal{B})$.

In the probabilistic treatment I assumed in Section 5 that the observations and the parameters could be defined on the same underlying probability space. In the present setting I assume that all statements regarding conceptual variables can be given as compatible propositions. The concept of an epistemic process is central in this book. Before any observations are made, all statements of the form $\theta = u$ are unknown, where θ is the relevant epistemic variable. After the observations are done, some proposition $A_k : (\theta = u_k)$ is known to some agent i in the simplest case. The statement that A_k is known to agent i may be written $K_i A_k$, and the statement that agent j knows that i knows A_k may be written $K_j K_i A_k$. A survey of the formal propositional logic related to such statements is given by Halpern (1995).